

School of Electrical Engineering and Computing
Department of Computing

Human Body Tracking and Pose Estimation
From Monocular Image Sequences

Yao Lu

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University

September 2013

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgement has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Abstract

Estimating human pose over time based on monocular view with no restriction on the human activities is an important field in computer vision due to its potentially wide applicability. There are many existing work in the field. Among them bottom-up approaches have been found to have big advantage in estimating large variety of human poses. They use independent body part detectors to find candidates for each body part, which are subsequently assembled together to estimate poses that match acceptable configurations of human posture based on the locations of the detections. Body part detectors can either be generic (shape-based) or specific (colour-based) to the appearance of the person being tracked. Bottom-up estimation does not require pre-established motion models and it works well with monocular view, unlike the top-down ‘predict-then-evaluate’ approaches. In order to achieve more accurate estimation, some recent successful bottom-up approaches have attempted to gain the benefits of both the generic and specific detectors by utilising generic detectors to bootstrap the learning of specific detectors. However, current approaches have some weaknesses in their selection of generic detections for building a specific colour appearance model. Specifically, some approaches rely on detecting at least one very distinctive posture in the video sequence using generic detectors to guarantee good training estimations for building the specific appearance model. Others avoid this by selecting the top few most likely generic detections, but must rely on the unverified assumption that those generic detection likelihoods are always accurate despite them being extracted from different data (different frames). In addition, when a specific appearance has been built for the final estimations, either the generic detections are discarded (thus ignoring shape-based information) or the generic likelihood maps are simply averaged with the specific appearance likelihoods. Furthermore, due to the relatively loose fit of body models, background pixels will also inevitably be included when training the specific appearance model. This will contaminate the specific appearance model constructed and reduce its effectiveness, yet no approaches have considered addressing this problem to date.

This thesis describes a bottom-up approach to estimate human pose over time based on monocular views with no restriction on the human activities. Firstly, this thesis shows that generic detection likelihood is not necessarily a representative of accuracy when comparing detections between frames, hence likelihood alone cannot be reliably used for selecting generic detections to train the specific appearance model. Instead, this thesis proposes to cluster generic pose estimations in terms of their colour to identify the subset of opti-

mal estimation that are accurate based on the more reasonable assumption that correct estimations will have a similar colour appearance since they will consistently track the same body part within an image sequence. The clustering approach for learning a specific appearance model has two advantages over existing approaches: (1) it eliminates the need of existing approaches for a distinct pose to exist in the sequence and uses multiple frames to build the colour model. This allows it to be applied to sequences with fewer restrictions on the postures or lighting conditions. (2) it avoids the need to rely solely upon the likelihood of the generic detections for training the specific appearance model by filtering out poor (though high-likelihood) detections using colour.

Secondly, the thesis proposes an effective method to ensure that the final estimations match both the generic and specific appearance models. Specifically, the final estimation is determined by filtering the top few most likely generic detections in a given frame based on the specific appearance model. Selecting the most suitable match from this filtered subset means that the final estimation is accepted by both the generic and specific appearance models. This has the advantage that evidences from both the generic shape-based and specific colour-based appearance model are satisfied in the final estimation decision, rather than discarding the generic detections or simply averaging over the two types of evidences.

Thirdly, the thesis proposes a method to identify and remove non-target (background) pixels from training samples used in learning a specific appearance model and build a less contaminated colour appearance model. This is implemented using clustering to group pixels of similar colour and separates target from non-target pixel clusters based on their distance to the central axis of the body part. The advantage of this method is that a colour appearance model built from a less contaminated colour profile is less likely to confuse background with the body part to be detected.

This thesis implements and evaluates a system that utilizes the three proposed methods to perform the task of estimating human pose over time based on monocular view with no restriction on the human activities. Experiments are conducted on several challenging, publicly-available video sequences and evaluated in terms of both overlap with the ground truth and body joint estimation error. Results demonstrate that the proposed system outperforms existing systems significantly, particularly for the most difficult body parts such as the lower arms and lower legs.

Acknowledgements

In many ways, this thesis could not have been possible without the contributions, great and small, direct and indirect, of many individuals over the course of the PhD.

Firstly, I would like to thank my supervisor, Professor Ling Li. She provided an unending source of motivation and support across the full spectrum, from giving me the opportunity to do a PhD and suggesting a really interesting topic to continuous guidance and help in all aspects of my research. In addition, I'd also like to thank my co-supervisor, Dr Patrick Peursum, who spent a lot of time discussing with me and give me a lot of inspiration.

Secondly, I would like to thank my parent who are always my strongest backing. In addition, I would like to thank my wife, Mrs Qing Tian. She gave me a lot of care in my daily life and stayed with me through the most lonely time.

I must also thank all the people who made their software available for free on the Web, providing invaluable tools for this research. These included Deva Ramanan at the University of California, Mykhaylo Andriluka at Max Planck Institute for Informatics. Finally, special thanks to the China Scholarship Council (CSC) and Curtin International Postgraduate Research Scholarships (CIPRS) for providing me with a scholarship without which I could not have continued to study.

Contents

Abstract	ii
Acknowledgements	iv
Publications	xiv
1 Introduction	1
1.1 Aims and Approach	3
1.1.1 Investigating Generic Detector Effectiveness	4
1.1.2 Learning Robust Colour Appearance Model	4
1.1.3 Reducing Contamination of the Appearance Model	5
1.2 Significance and Contributions	6
1.2.1 Investigating the Relationship between Accuracy and Likelihood	6
1.2.2 Clustering for Learning a Robust Specific Appearance Model	7
1.2.3 Estimations from Both Generic and Specific Appearance Models	7
1.2.4 Reduction of Pixel Contamination	8
1.3 Structure of the Thesis	8
2 Related Work	10
2.1 Human Detection and Tracking	12
2.1.1 Background Subtraction	12
2.1.2 Motion-based Approaches	16
2.1.3 Appearance-based Approaches	17
2.1.4 Shape-based Approaches	19
2.1.5 Section Summary	21
2.2 Pose Estimation: Model-free Approaches	21
2.2.1 Learning-based Approaches	21
2.2.2 Example-based Approaches	23
2.2.3 Section Summary	23
2.3 Model-based Pose Estimation: Top-down Approaches	24
2.3.1 Human Model	25
2.3.2 Likelihood Function	27
2.3.3 Dynamic Model	29
2.3.4 Posterior Estimation	29
2.3.5 Section Summary	31
2.4 Model-based Pose Estimation: Bottom-up Approaches	31

2.4.1	Segmentation-based	31
2.4.2	Pictorial Structure-based	32
2.4.3	Section Summary	37
2.5	Pictorial Structure Using a Specific Appearance Model	37
2.5.1	Approaches Review	38
2.5.2	Specific Appearance Model vs Generic Appearance Model	39
2.5.3	Weakness of the Approaches based on Specific Appearance Model	41
2.5.4	Section Summary	43
2.6	Chapter Summary	43
3	Pictorial Structure Framework	45
3.1	Pictorial Structure Framework	46
3.1.1	Statistical Framework of Pictorial Structure Model	48
3.1.2	Algorithm for Finding the Optimal Match	50
3.1.3	Algorithm for Sampling from the Posterior	53
3.2	Generic Human Pose Detector	56
3.2.1	The Configuration Prior	57
3.2.2	The Appearance Likelihood Model	58
3.2.3	Outputs of the Generic Pose Detector	59
3.3	Analysing the Assumption of Ferrari	61
3.4	Investigating the Generic Detector	62
3.4.1	Analysis of the Optimal Estimations	63
3.4.2	Multiple Probable Alternative Estimations	66
3.5	Chapter Summary	69
4	Combining Generic and Specific Appearance Model	71
4.1	System Overview	73
4.2	Building a Specific Appearance Model	74
4.2.1	Feature Extraction and Representation	76
4.2.2	Mean Shift for Clustering	77
4.2.3	Analysis of the Clusters Obtained by Mean Shift	79
4.2.4	Specific Appearance Model	81
4.3	Tracking	81
4.3.1	Fusing Generic and Specific Appearance Detections	83
4.3.2	Local Search	85
4.4	Experiments and Discussion	89
4.4.1	Evaluation and Metrics	90
4.4.2	Experimental Results – Walking	92
4.4.3	Experimental Results – Non-walking Sequences	100
4.5	Chapter Summary	101

5	Building an Uncontaminated Specific Appearance Model	103
5.1	System Overview	104
5.2	Recap of Prerequisite Approaches	106
5.3	Removing the Non-target Pixels from Contaminated Correct Estimations	106
5.3.1	Basis of Removing Non-target Pixels	108
5.3.2	Extracting Target Pixels via Pixel Analysis	113
5.4	Building Target-pixel Classifiers and Labelling Target Pixels	117
5.4.1	Building Target-pixel Classifiers	118
5.4.2	Applying Target-pixel Classifiers to Each Frame	119
5.5	The Uncontaminated Appearance Model for Tracking	122
5.5.1	Specific Appearance Template	122
5.5.2	Appearance Likelihood Model	123
5.5.3	Tracking	124
5.6	Experiments and Discussion	125
5.6.1	Baseball and Walking Sequences	126
5.6.2	Combo Sequence	130
5.7	Chapter Summary	134
6	Conclusions	136
6.1	Summary of Contributions	140
6.2	Future Work	141
6.2.1	Scale Changes	141
6.2.2	Motion Blur	142
6.2.3	Multiple Targets	142
6.2.4	Efficiency	143

List of Figures

2.1	Human shape model with kinematical model. (a) 2D model (reprinted from Huang and Huang (2002)); (b) 3D volumetric model consisting of superquadrics (reprinted from Kehl and Gool (2006)); (c) 3D surface model (reprinted from Carranza <i>et al.</i> (2003)).	25
2.2	3D kinematical model of Sidenbladh <i>et al.</i> (2000)	26
2.3	Configurations of the pixel map sampling points $p_i(X_t, I_t)$ for the edge based measurements (a) and the foreground segmentation measurements (b).	28
2.4	A rectangular body part. Area 1 is the central area inside the part, and Area 2 is the border area around Area 1.	34
2.5	(a) The log-polar grid shows 9 location bins used in shape context. (b) 12 location bins used in the variant of shape context.	35
2.6	(a) Tree model capturing the kinematic model of human body.(b) 12 location bins used in the variant of shape context. (Tor=Torso, LUA=Left-Upper-Arm, RUA=Right-Upper-Arm, LUL=Left-Upper-Leg, RUL=Right-Upper-Leg, LLA=Left-Lower-Arm, RLA=Right-Lower-Arm, LLL=Left-Lower-Leg and RLL=Right-Lower-Leg)	36
2.7	An example from Ramanan <i>et al.</i> (2007) is used to demonstrate that in comparison with a generic appearance model a specific appearance model facilitates the removal of interference from background clutter in estimating human pose.	40
2.8	To describe the drawbacks of approaches only based on colour appearance model, some parts of Ramanan's results are presented. Since the upper arms has similar colour with the torso and the lower arms are hidden by the torso, the system is unable to detect where the upper arms are.	41
2.9	Figure (a) provides several inaccurate estimations for torso orientation from Ramanan <i>et al.</i> (2007). Figure (b) provides a demonstration on why a specific appearance model confuses the orientation estimation.	42
2.10	Some examples are provided to show that a specific appearance model can result in wrong estimations when two limbs are not well separated.	43
3.1	Kinematic tree model. Each node represents a body part. The root node is denoted by l_{torso} . The subscripts denote the body parts. In the abbreviation of subscripts, the first character l or r denotes right or left. The second character u or l denotes upper or lower. The last character a or l denotes arm or leg. The dependency relations are denoted by directed edges.	57

3.2	An analysis of the correlation between detection accuracy and the likelihood score of the optimal estimations in different frames is conducted, which is based on the Baseball video sequence from Ramanan <i>et al.</i> (2007). Each bar represents the likelihood of the optimal estimation in a single frame, sorted on likelihood with intensity indicating ground-truth accuracy. (a) right lower arm. (b) left lower arm. Note that likelihood (order) and accuracy (intensity) have very little correlation.	61
3.3	(a) The baseball sequence from Ramanan <i>et al.</i> (2007). (b) The badminton sequence from online video.	62
3.4	Several examples demonstrate that both the correct and incorrect estimations possibly exists within the optimal estimations	64
3.5	Several examples demonstrate that both the correct and incorrect estimations are possibly chosen as the optimal estimations	65
3.6	The percentage of frames where the optimal estimations are correct for each body part on Baseball and Badminton sequences.	66
3.7	Two examples demonstrating why multiple probable alternative estimations are needed. When the optimal estimations are incorrect as shown in Figure (a) and (c), the correct estimations are highly likely to exist in the set of multiple probable alternative estimations as shown in Figure (b) and (d). .	67
3.8	Percentage of frames where the correct (ground-truth) limb estimations exist within the sampled estimations. Note that the percentage of frames no longer increases after the number of samples reaches 15. This is because that the correct limb estimations are impossibility to be detected when they are invisible in some cases such as severe motion-blur and self-occlusion. In such cases, the percentage of correct estimation is unable to reach 100% no matter how many possible sampled estimations are checked.	68
3.9	Percentage of frames where the correct (ground-truth) limb estimations exist within the sampled estimations on Badminton sequence.	68
4.1	The overview of our approach	73
4.2	(a) The accurate torso estimations from the preliminary estimations have similar colour. (b) The inaccurate torso estimations from the preliminary estimations will have different colours with accurate estimations.	75
4.3	A preliminary estimation for body part m in frame t is denoted by $\{x_m^t, y_m^t, \theta_m^t\}$, which corresponds to a bounding box. The center and orientation of the bounding box is respectively denoted by $\{x_m^t, y_m^t\}$ and θ_m^t	76
4.4	The analysis for the results of clustering.	79

4.5	Some examples to compare the estimations from Ramanan <i>et al.</i> (2007)'s colour-only detector with the estimations from the generic detector described in Chapter 3.	82
4.6	Process of filtering. Symbols of solid circle represent the elements in the set U_m and other symbols represent the elements in the set S_m . The elements locating in the main cluster are selected into the set U_m^*	85
4.7	The endpoints for all body parts in the human model.	87
4.8	The endpoints for all body parts in the human model.	88
4.9	Three sequences are used to compare our system with Ramanan's and Ferreri-Core in this chapter. Walking is short for HE1_S2_Walking_1_C1. Throwing is short for HE1_S4_Throw_Catch_2_C1.	90
4.10	The screenshot of tracking results in Baseball and Walking sequences from Ramanan <i>et al.</i> (2007)'s system and proposed system	93
4.11	Evaluating three systems' tracking performance on Baseball and Walking sequences based on Metric-1.	94
4.12	Performance in tracking Torso for Baseball and Walking Sequence is evaluated by the Metric-2.	96
4.13	Scheme difference between the proposed system and Ramanan's system. . .	97
4.14	The evaluation results based on Metric-2 for three systems in Baseball sequence and Walking sequence. Error is average error across all body parts.	99
4.15	Screenshot of tracking results on Throwing sequences	100
4.16	The evaluation results based on Metric-1 for two systems (Ferrari-Core and proposed) in Throwing sequence that contains no lateral walking pose. . . .	100
4.17	The evaluation results based on Metric-2 on two systems (Ferrari-Core and proposed) in Throwing sequence.	101
5.1	(a) Overview of the improved human pose tracking system (b) Visualization of how we build an uncontaminated specific appearance model for human pose tracking based on the results from generic human pose detection and the results of colour histogram clustering. The process of training the torso appearance is used to illustrate our approach.	105
5.2	Two examples show that both target pixels and non-target pixels co-exist in correct estimations. Figure (a) shows that non-target pixels may come from the background. Figure (b) shows that non-target pixels may also come from the non-target body parts.	107
5.3	Walking sequence (HumanEva_I_Walking_S2) from HumanEva dataset (Sigal and Black, 2006).	109

5.4	The comparison in the quantities of target pixels and non-target pixels specified by contaminated correct estimations for each body part in the Walking sequence.	110
5.5	(a) An estimation is represented by a bounding box in an image. The green area represents the central area and the white area represents the border area of the bounding box. (b) A figure to demonstrate the columns of a bounding box.	111
5.6	The percentage of target pixels and non-target pixels for each column of bounding box across all correct estimations in body parts LLL, LUL, LLA, RUA, Head and Torso.	112
5.7	Combo sequence (Human_Eva.I_S2_Combo_2_C2) from HumanEva dataset (Sigal and Black, 2006).	114
5.8	Two examples for pixels clustering. Figure (a) is the result of pixels clustering for left lower leg (single-colour body part). Figure (b) is the result of pixels clustering for left upper arm (multiple-colour body part).	116
5.9	A demonstrating figure of image patch $a_m^{t_n}$ to explain how to compute a perpendicular distance d of the pixel (x, y) with respect to the vertical axis of the bounding box.	116
5.10	Several examples of marking target pixels for single-coloured body part using learned Gaussian appearance classifiers. The frames shown in this figure are representative and typical in Combo sequence (Human_Eva.I_S2_Combo_2_C2) from HumanEva dataset (Sigal and Black, 2006).	120
5.11	Several examples of marking target pixels for multiple-coloured body part using learned Gaussian appearance classifiers.	121
5.12	(a) Each body part in the human model relates to a specific appearance template where green represents the target-pixel area and red represents the non-target-pixel area. (b) Appearance template for an individual body part.	123
5.13	Three sequences are used to compare our proposed system with three other systems (ie: Ramanan's system, Ferrari-Core system, and CLUSTERING system). Walking is short for HE1_S2_Walking_1_C1. Combo is short for HE1_S2_Combo_2_C2.	126
5.14	The performance comparison between Clustering and Classifier systems based on Metric-1 in Baseball and Walking Sequences.	127
5.15	The screenshot of tracking results in Baseball and Walking sequences from CLUSTER (CLS) and CLASSIFIER (CLF).	128
5.16	The performance comparison between Clustering and Classifier systems based on Metric-2 in Baseball, Walking, Throwing and Combo sequences.	129

5.17	The screenshot of tracking results in Combo sequences from CLUSTER (CLS) and CLASSIFIER (CLF).	130
5.18	The errors in estimating the configuration of head and torso in every frame of Combo sequence by four systems.	131
5.19	The errors in estimating the configuration of LLA and LUA in every frame of Combo sequence by four systems.	132
5.20	The errors in estimating the configuration of LUL and LLL in every frame of Combo sequence by four systems.	133

List of Tables

5.1	Average distance to central axis for target pixels, all pixels and non-target pixels specified by correct estimations.	113
-----	--------------------------------------------------------------------------------------------------------------------------------	-----

Publications

This thesis is based upon several works that have been published (or submitted) over the course of the authors PhD, listed as follows in chronological order:

- Yao Lu, Ling Li and Patrick Peursum (2012). Background Suppression for Building Accurate Appearance Models in Human Motion Tracking. *Proceedings of International Conference on Digital Image Computing Techniques and Applications*.
- Lu Yao, Ling Li, Patrick Peursum (2012). Human Pose Tracking Based on Both Generic and Specific Appearance Models. *Proceedings of the International Conference on Control, Automation, Robotics and Vision*.
- (under review) Lu Yao, Ling Li, Patrick Peursum (2012). Tracking people combining both specific and generic appearances. *Journal of the Computer Vision and Image Understanding*.

Chapter 1

Introduction

Human pose estimation has been the focus of considerable research effort over the past few decades due to its extensive applicability in computer vision. Human pose estimation is the process in which the configuration of human body parts is estimated from various types of input. In some approaches pose estimation depends on electromagnetic sensors that are attached to the human body, but these are relatively costly and sometimes unstable systems. Since many applications (such as visual surveillance, human-computer interaction, 3D animation and robotic control) would benefit from cheaper and more convenient vision-based human pose estimation using cameras, this topic has received increasing attention in recent years.

One of the major difficulties in vision-based human pose estimation is the high number of degree-of-freedom (DOF) in the human body's movement that must be estimated. This difficulty is compounded by the problem of self-occlusion, where body parts occlude each other. In addition, varying lighting conditions that affect appearance can also hinder accurate estimation. Finally, additional complications arise in cases where camera parameters are unknown, such as the location of the camera in comparison to the human. Due to the technical challenge and the significant potential of human pose estimation there exist numerous works on the topic in the literature. However, in practice the resulting systems typically require some limiting assumption such as multiple cameras, known camera information, specified activities or manual initialization.

The typical system of vision-based human pose estimation involves a model-based approach, in which an observation is captured and provided as input to the model to obtain pose estimations. Such model-based approaches can be categorized as either top-down or bottom-up. Top-down approaches produce a set of hypotheses for pose and evaluate how well each of these hypotheses match the observed image(s) to arrive at a pose estimation. This kind of approaches usually requires either observations from multiple cameras or an effective motion model trained for a specific human movement. In contrast, bottom-up approaches use independent detectors to find candidates for each body part, then estimates pose by finding assemblies of body parts that match acceptable configuration(s) of human posture based on the locations of the detections. This kind of approaches does not require

motion models and works well with single view, although it is also computationally more expensive than a top-down approach. In this thesis, the research goal is to accurately estimate human pose over time based on monocular view with no restriction on the human activities and therefore is focused on the bottom-up approach.

A popular bottom-up approach is to use the pictorial structure (Felzenszwalb and Huttenlocher, 2005) to probabilistically encode the spatial relations between body parts and transform a set of limb detections into a distribution of valid pose configurations and select the optimal pose as the estimation. Limb detectors themselves vary, with generic detectors that encode the generic appearance of body parts (such as shape or edges) which can be used to model any human but can become confused with the background. In contrast, specific detectors that model the characteristics of a particular tracked human (often in terms of colour) is typically more accurate but must be trained for the specific human. Recent successful approaches (Ramanan *et al.*, 2007; Ferrari *et al.*, 2009) have attempted to gain the benefits of both by utilising generic detectors to bootstrap the learning of specific detectors for subsequent accurate tracking.

Ramanan *et al.* (2007) utilises what they refer to as a stylized pose detector, which is a generic human pose detector designed to find distinct poses such as a lateral walking posture. The stylized pose detector is used to estimate the human pose in a frame where a lateral walking pose appears, and from the best few detections, selects a single example (the median) to learn a specific (colour-based) appearance model. This specific appearance model is then encoded in a pictorial structure to estimate human pose in each frame. Accuracy is evaluated by defining a correct detection as one whose limb estimate overlaps the ground truth by at least 50%. However, the specific appearance model cannot be learned when the lateral walking pose does not exist in the video sequence, and the reliance on using a single frame to learn the colour appearance means the specific appearance model may not be robust to lighting variations throughout the video sequence.

Unlike Ramanan *et al.* (2007), Ferrari *et al.* (2009) uses generic body part detectors to estimate arbitrary poses in all frames, selecting the optimal estimation from each frame calculated according to the pictorial structure likelihood. Only the top few of these per-frame optimal estimations are subsequently used to learn the specific appearance model for the tracked human, under the assumption that these top few will be the most accurate. A second pose estimation phase is then executed by applying the specific appearance model and pictorial structure to all frames, followed by averaging the two probability maps (generic and specific) to extract a final estimation per frame. The advantages of this method are that it does not require a stylized pose to appear, that it utilizes evidence from both generic and specific detectors, and it samples colours from multiple frames

and hence includes lighting variations into the specific appearance model. However, the effectiveness of the specific appearance model depends on the assumption that generic detection likelihood is representative of accuracy *between* frames and so that the top few optimal estimations will be the most accurate. This assumption was not verified by Ferrari *et al.* (2009) – the issue here is that it is comparing pose estimations that have been extracted from different data (different frames), hence it is not certain that detection accuracy and detection likelihood will correlate *between* frames. Moreover, their choice to merge the generic and specific appearance maps via averaging does not ensure that both models must agree on the final estimation. For example, a very high likelihood in colour and low likelihood in generic can still produce a acceptable estimation in their method.

In addition to the aforementioned issues, it is inevitable that the optimal estimations from the generic human pose detector will not always be guaranteed to be accurate - occasionally they are incorrect for some frames. However, both Ramanan *et al.* (2007) and Ferrari *et al.* (2009) directly use the optimal estimations to learn a specific appearance model without an attempt to handle such a possibility beyond assuming that the most likely set of optimal estimations are accurate. Unfortunately, including inaccurate estimations in the training samples will ‘contaminate’ any specific appearance model built. Furthermore, any estimation (even a so-called *accurate* estimation) might not perfectly overlap its corresponding body part. Background pixels will also inevitably be included, further contaminating the specific appearance model constructed, and consequently affect the accuracy of the pose estimations.

1.1 Aims and Approach

The specific aims of the thesis are as follows:

1. Quantitatively investigate the effectiveness of generic (edge-based) body part detectors based on the pictorial structure, with the objective of assessing the validity of the assumptions made by Ferrari *et al.* (2009):
 - (a) Analyze whether the optimal estimation per frame from a generic detector is in fact consistently accurate.
 - (b) If not, explore whether accurate configurations consistently exist within the top N most likely estimations in a given frame and thus could potentially be utilised.

2. Learn a robust specific (colour) appearance model that produces an effective model for human pose tracking despite inaccurate estimations existing in the optimal estimations.
3. Develop a human pose detector whose final estimation conforms to the evidence of *both* the generic and specific appearance models by filtering the top N generic estimations based on their colour.
4. To analyse the issue of contamination of the colour appearance model with background pixels, and develop algorithms to reduce this contamination in order to improve the effectiveness of a colour model for tracking.

1.1.1 Investigating Generic Detector Effectiveness

In order to learn a specific appearance model, [Ferrari *et al.* \(2009\)](#) makes two critical assumptions: 1) the optimal (most likely) generic estimation in a frame is the most accurate choice for that frame; and 2) the likelihood of generic estimations between different frames are comparable and can represent their relative accuracy. Selecting the top few optimal estimations by comparing likelihood of each frame’s optimal estimation should then represent the most accurate set of generic detections in the video sequence to learn a specific appearance model. However, these assumptions were not empirically shown to be the case, hence it is necessary for this thesis to first examine their validity in some representative video sequences. This examination is composed of three parts:

- Examine the accuracy of the optimal-likelihood generic estimation in each frame to establish the degree to which the optimal estimations can be relied upon.
- Analyze the top N most likely estimations of each frame to investigate the relationship between likelihood and accuracy in detections within a single frame.
- Analyze the relative accuracy of the set of optimal generic estimations between frames to determine whether estimations from different frames can be reasonably compared using likelihood.

1.1.2 Learning Robust Colour Appearance Model

Both [Ramanan *et al.* \(2007\)](#) and [Ferrari *et al.* \(2009\)](#) learn a specific appearance model based on the optimal estimations from a generic human pose detector, but both must

assume that the optimal estimations used are accurate. In difficult video sequences such an assumption cannot be guaranteed and training samples will include inaccurate estimations, leading to learning a less effective specific (colour) appearance. Thus to address Aim 2, this thesis proposes to cluster generic pose estimations in terms of their *colour* to identify the subset of optimal estimations that are accurate, based on the assumption that correct estimations have a similar colour appearance. A specific appearance model is then learned using these identified correct optimal estimations. For frames whose optimal estimations do not cluster as ‘correct’, the top N most likely generic estimations for each incorrect frame is filtered to find candidate postures that conform to the correct cluster. This approach obtains pose estimations that agree with the evidence of both the generic and specific appearance models, thus addressing Aim 3. Note that although [Ramanan *et al.* \(2007\)](#) also utilises clustering of colour, they do so independently for each individual part detector simply as a means of selecting a single example (a median) for representing the part in the specific appearance model. In contrast, this thesis utilises clustering as a central part of the estimation process, to identify accurate generic optimal estimations and further search for accurate (though sub-optimal in terms of likelihood) detections.

1.1.3 Reducing Contamination of the Appearance Model

As discussed, [Ramanan *et al.* \(2007\)](#) and [Ferrari *et al.* \(2009\)](#) learn a specific appearance model based on estimations from a generic human pose detector. Since the precise shape, size and boundaries of the body parts are unknown, any generic estimation (even if accurate) generally cannot perfectly overlap with its corresponding body part. This means that non-target colours from background pixels or pixels from other body parts will unavoidably be included in the process of learning the specific appearance model, thus contaminating the specific appearance model. To address Aim 4, we seek to separate target and non-target pixels in a contaminated estimation of a body part based on an analysis of their colours and pixel locations. Specifically, it is hypothesized that target pixels are more likely to appear in the central area of a detection bounding box for a body part, a reasonable assumption if detection is accurate. Due to the consistency of a body part’s appearance in a video sequence and the part’s difference in colour from non-target pixels, it is proposed that clustering pixel colours will separate target pixels into one set of clusters and non-target pixels into other clusters. Identifying target clusters is then a matter of identifying clusters that are largely produced by pixels that are more centrally located in the bounding box.

1.2 Significance and Contributions

This thesis makes four main contributions to the field of computer vision — (1) a quantitative analysis of the relationship between accuracy and likelihood of human pose estimations based on generic detections; (2) the proposal of clustering based on colour as a means of separating accurate from inaccurate generic pose estimations in cases where detection likelihood is not a reliable indicator of accuracy; (3) the enforcement that final pose estimations concur with the evidence from both generic and specific appearance models by utilising the colour-based clustering to identify and select an accurate pose from the top N shape-based pose estimations of each frame; and (4) a method to reduce background colour contamination in a specific appearance model by use of pixel clustering and pixel location analysis to construct a new and more effective specific appearance model.

1.2.1 Investigating the Relationship between Accuracy and Likelihood

The first major contribution of this thesis is a quantitative and empirical investigation of the relationship between the accuracy and likelihood of estimations from generic human pose detection. In order to examine the validity of the assumptions made by [Ferrari *et al.* \(2009\)](#), two representative sequences are processed with a generic pose detector and this thesis seeks to answer three questions:

1. What proportion of optimal per-frame estimations are accurate estimations?
2. Does an accurate estimation consistently exist within the top N estimations of each frame?
3. Is detection likelihood a useful measure for comparing the accuracy of optimal estimations between frames despite being based on different data?

The answers to these questions are critical in clearly defining the requirements that a pose estimation system must address in order to achieve robust and accurate tracking. It is expected that whilst generic detectors will often be accurate, a system cannot rely on all (or even most) optimal estimations being accurate. However, we contend that it is less fragile to assume that an accurate pose estimation will usually exist somewhere within the top N estimations of a frame, with the challenge being how to identify this accurate estimation. Finally, we hypothesize that, in contrast to Ferrari’s assumption, the likelihood obtained from different frames is not usefully comparable, and so can lead to selection of training samples that are not robust.

1.2.2 Clustering for Learning a Robust Specific Appearance Model

The second major contribution of this thesis is the use of clustering to learn a robust specific appearance model. Unlike existing approaches (Ramanan *et al.*, 2007; Ferrari *et al.*, 2009), the proposed approach explicitly takes the view that a generic detector’s likelihood is not always (or even mostly) a reliable indicator of accurate detections for training a more powerful specific appearance model. Instead, it is recognized that whilst many detections will be accurate, many others will be inaccurate or false positives from the background. Thus the approach is to identify accurate detections based on the reasonable assumption that accurate detections will have consistency in their body part colour features. Since colour is not a factor in generic shape-based detection, the colour features of inaccurate detections will differ from the correct model and even from each other. Clustering based on colour will then group similarly-coloured detections together, with the largest cluster expected to include the accurate detections.

The significance of this is the inclusion of alternative evidence in the selection process, namely colour. This eliminates the need of Ramanan *et al.* (2007) for a special stylized pose to exist in the sequence and will use multiple frames to build the colour model, so can be applied to sequences with fewer restrictions on the postures or lighting conditions found in the sequence. It also avoids the need to rely solely upon the likelihood of the generic detections as Ferrari *et al.* (2009) do, filtering out poor (though high-likelihood) detections for training the specific appearance model.

1.2.3 Estimations from Both Generic and Specific Appearance Models

The third major contribution of this thesis is an effective method to make the final estimations conform to both the generic and specific appearance models. The final estimation is determined by filtering the top few most likely generic estimations based on the specific appearance model. Selecting the most suitable match from this filtered subset means that the final detection is accepted by both the generic and specific appearance models.

It has the advantage that evidences from both the generic shape-based and specific colour-based information are satisfied in the final estimation decision. This is in contrast to Ramanan *et al.* (2007), who discards the generic detection evidence once the specific appearance model is built from it, and Ferrari *et al.* (2009), who simply averages the two types of evidences.

1.2.4 Reduction of Pixel Contamination

The final major contribution of this thesis is the analysis of the characteristics of contamination in accurate estimations, and subsequent proposal of a method to identify and remove non-target (background) pixels from training samples used in learning a specific appearance model. The goal is to learn an uncontaminated (or less contaminated) specific colour appearance model in order to improve tracking effectiveness. To our knowledge, little work has been done on this problem beyond assuming that the body part detector bounding boxes fit tightly enough to be acceptable (Ramanan *et al.*, 2007; Ferrari *et al.*, 2009), or to build complex body models that are perfectly fitting the specific person’s body Balan *et al.* (2005). We suggest that target pixels are more likely to appear in the central area of a contaminated accurate estimation. Thus we propose to use clustering to group pixels of similar colours and utilise each group’s distance to the central axis to identify target (true body part) pixel colours.

The significance of such an approach is that a colour appearance model built from a less-contaminated colour profile is less likely to confuse background with the body part to be detected. This would be expected to result in better tracking performance than a colour model that was built with non-target contamination.

1.3 Structure of the Thesis

This thesis is organized as follows. In Chapter 2, a review of related work in the fields of human finding and tracking and human pose estimation is presented. The existing approaches for finding and tracking human within a scene are first briefly explored. This is followed by discussion on model-free approaches used for human pose estimation and its limitations. Then model-based approaches for human pose estimation, which can be divided into top-down approaches and bottom-up approaches, are reviewed respectively. In top-down approaches, some typical human models and estimations approaches are explored and their limitations are analyzed. Previous work in bottom-up approaches are then discussed, focusing on the approaches under pictorial structure framework. Finally, a detailed analysis is conducted in the approaches (Ramanan *et al.*, 2007; Ferrari *et al.*, 2009) that are closely related to the research presented in this thesis.

In Chapter 3, the framework of the pictorial structure and some relevant algorithms are first reviewed. Then a generic human pose detector is built based on the framework of pictorial structure. Finally, the characteristics of the generic human pose detector are

investigated by verifying the assumption of [Ferrari *et al.* \(2009\)](#) and analyzing the accuracy of the optimal estimations and sub-optimal estimations.

In Chapter [4](#), a tracking system for estimating 2D human pose over time is implemented. In this system, an approach is first proposed to automatically learn a specific (colour-based) appearance model using clustering based on the optimal estimations from generic human pose detector. Another approach is proposed to obtain the final estimations that satisfy both the generic and specific appearance models. Experiments are conducted to demonstrate that the proposed system outperforms the systems proposed by [Ramanan *et al.* \(2007\)](#) and [Ferrari *et al.* \(2009\)](#).

In Chapter [5](#), the problem of background pixels contaminating the specific appearance model is addressed. In order to learn an uncontaminated specific appearance model, an approach is proposed to identify and remove non-target (background) pixels from the optimal estimations of the generic human pose detector. An uncontaminated specific appearance model is hence built to be applied in a tracking system. Experiments are conducted to test the performance of our tracking system that uses the uncontaminated specific appearance model in comparison with [Ramanan *et al.* \(2007\)](#), the core of [Ferrari *et al.* \(2009\)](#) and the tracking system proposed in Chapter [4](#).

Finally, Chapter [6](#) provides a summary of the thesis, its contributions and potential future directions.

Chapter 2

Related Work

The research goal in this thesis is to accurately estimate 2D human configurations over time based on monocular view. There is no restriction on the human activities and the camera parameters are unknown. To achieve this goal, the scope of research in this thesis involves elements from two branches of human motion analysis in computer vision, i.e., human finding and tracking, and human pose estimation. Human finding and tracking, in which the entire human is detected and tracked as a single object, is a fundamental requirement for most computer vision systems for estimating human pose. In most cases, a system of pose estimation requires knowing a rough region that contains a human before a pose can be estimated. Thus human finding and tracking is usually a pre-processing step for human pose estimation. Moreover, many basic techniques and approaches for estimating human pose are developed from the techniques and approaches for human finding and tracking. Therefore, in this chapter, the basic approaches for human finding and tracking are reviewed first for better understanding of the approaches to human pose estimation. In contrast to human finding and tracking, human pose estimation focuses on estimating the location of each body part. There is a wide variety of approaches for human pose estimation designed for different purposes and depending on different assumptions. In this chapter, we will review some classic approaches and analyze their advantages and disadvantages.

Human finding and tracking usually consists of two steps: 1) finding (detecting) a human in each frame of a sequence; and 2) building temporal correspondences between these detections. The approaches for finding (detecting) a human in a single frame can be separated into four categories: background subtraction, motion-based detection, appearance-based detection and shape-based detection. Background subtraction is originally designed for detecting foreground objects rather than just detecting humans. Due to its simplicity and effectiveness, this type of method is widely used for detecting a human in a single frame or as a pre-processing step for other human detection approaches. Traditional approaches for background subtraction are to detect the foreground objects as the difference between the current frame and an image of the scene's static background. Since the changes in illumination and background etc. are not taken into consideration, these approaches can only be applied in controlled indoor environments. In 1999, [Stauffer and Grimson \(1999\)](#)

presented the idea of representing each pixel by a mixture of Gaussians (MoG) and updating each pixel with new Gaussians during run-time. This allows background subtraction to be used in outdoor environments. In this chapter, the approaches for background subtraction are reviewed separately in four aspects, including background representation, classification, background updating, and background initialization. Unlike background subtraction, the other three approaches are specifically designed for human detection but using different ideas. Motion-based detection approaches are based on the idea that the differences in consecutive frames arise from the moving human, which provides ways to detect the human by finding the motion. The motion is measured using either optical flow or image differences. Appearance-based detection approaches are based on the ideas that: 1) the appearance of human and background is different; and 2) different individuals have different appearances. Finally, the shape-based detection approaches are built on the idea that the shape of a human is often very different from the shape of other objects in a scene.

Given the location of a human from a detection method, human pose estimation can proceed. The approaches for human pose estimation can be broadly divided into discriminative and generative approaches. The discriminative approaches, also called model-free approaches, encompass approaches which build a direct relationship between image observations and human postures rather than matching a pre-defined human model to image observations. In contrast to model-free approaches, the generative (model-based) approaches first define a human model (modeling) and then match the pre-defined model to image observations (estimation). Model-based approaches can be divided into top-down approaches and bottom-up approaches. The top-down approaches can also be called model-based analysis-by-synthesis approaches. This type of approaches first builds a human model which consists of an explicit geometric representation of human shape and its kinematic structure. Human pose is estimated by optimizing the similarity between a model projection and the observed images. The bottom-up approaches also need to first build a human model in which the appearance model of each body part and the connection relations between them are defined. Human pose is then estimated in three steps: 1) finding candidates for each body part from image observations based on pre-defined appearance models; 2) assembling these candidates and computing the matching degree of each assembly according to the human model; and 3) selecting the assembly with the maximum matching degree as the final estimation.

This chapter is organized as follows. Section 2.1 describes the existing approaches for finding and tracking human within a scene, especially for the approaches closely related to human pose estimation. This is followed by a discussion in Section 2.2 on model-free approaches used for human pose estimation. At the end of this section, we will analyze the

limitations of model-free approaches and discuss why model-based approaches are more suitable than model-free approaches for the objective in this thesis. Top-down approaches for human pose estimation are discussed in Section 2.3 where we will describe some typical human models and estimation approaches, and analyze the limitations of these approaches. Section 2.4 discusses previous work in bottom-up approaches, focusing on the approaches under pictorial structure framework. In Section 2.5, we will discuss the approaches that are closely related to the research presented in this thesis. Finally, a summary of the chapter is presented in Section 2.6.

2.1 Human Detection and Tracking

Human detection and tracking is for detecting the entire human body and tracking it as a single object. This is often a preprocessing step for human pose estimation. We categorize the approaches in accordance with the type of image measurements: appearance-based approaches, shape-based approaches, and motion-based approaches. Background subtraction as a classic object detection and tracking approach is first discussed.

2.1.1 Background Subtraction

Background subtraction was used as a powerful tool in controlled indoor environments until the late 90s. In 1999, to explore applications of background subtraction in outdoor environments, Stauffer and Grimson (1999) proposed the idea of representing each pixel by a mixture of Gaussians (MoG) and updating each pixel with new Gaussians during run-time. Slow changes in a scene can be modelled by recursive updating but rapid changes cannot be modelled. Although the method proposed by Stauffer and Grimson (1999) has become the standard of background subtraction, a lot of advances have been seen in background representation, classification, updating, and initialization since then.

2.1.1.1 Background Representation

The mixture of Gaussians (MoG) representations are originally applied in the RGB colour space, but other colour spaces have also been explored (Kristensen *et al.*, 2006). To detect shadow-pixels wrongly classified as object-pixels (Prati *et al.*, 2003; Han and Davis, 2012), a representation is often applied in a colour space where the colour and intensities are

separated, such as the YUV (Wren *et al.*, 1997), HSV (Cucchiara *et al.*, 2003) and the normalized RGB (McKenna *et al.*, 2000) colour space. A MoG can be used in a 3D colour space to correspond to ellipsoids or spheres (depending on the assumptions on the covariance matrix) of the Gaussian representations (McKenna *et al.*, 2000; Stauffer and Grimson, 1999; Zhao and Nevatia, 2004b). Other geometric representations include truncated cylinders (Kim *et al.*, 2005) and truncated cones (Fehl *et al.*, 2006).

Representations using different concepts have been developed. In Elgammal *et al.* (2000), a kernel-based approach is developed where a background pixel is represented by the individual pixels of the last N frames. Haritaoglu *et al.* (2000) proposed to represent background pixels by their minimum and maximum value together with the maximum change allowed of the value in two consecutive frames. In Heikkilä and Pietikainen (2006), a background pixel is represented using a bit sequence. Each bit carries the information of whether the value of a neighbouring pixel is more than the value of the pixel of interest. Such a representation is also called a texture operator. It makes the background model invariant to monotonic illumination changes. A pixel’s neighbours are also used by Oliver *et al.* (2000); Jiang *et al.* (2012) to represent the background pixels. An eigenspace representation is used to represent the background and new objects are detected by comparing the input image and a reconstructed image.

Eng *et al.* (2003, 2006); Kim *et al.* (2012) proposed a block-based background modelling. A background is divided into a number of non-overlapping blocks and the background model is learned over time. Depending on homogeneity, the pixels in each block are categorized into at most three classes. The mean values of these classes in a block are applied to represent the background for this block. This representation is known as a spatio-temporal representation. In Heikkilä *et al.* (2004), texture operators are used in a spatio-temporal block-based (overlapping blocks) background segmentations. In Monnet *et al.* (2003) and Zhong and Sclaroff (2003), the spatio-temporal representation is also used when a predicted region representing the background is found by an autoregressive process.

When choosing a representation for the background, not only the accuracy but also the speed of the implementation and the application need to be considered. In practice, the overall accuracy of background subtraction is not only determined by the representation but also the classification, updating and initialization. In some cases, due to the requirement on the application speed, a simple representation (Cucchiara *et al.*, 2003) is chosen for the background but good results can still be obtained due to the advanced classification and updating.

The MoG representation is by far the most widely used representation for background but this representation does not suffice for scenes with dynamic backgrounds. For methods directly aimed at dynamic backgrounds refer to [Monnet *et al.* \(2003\)](#); [Zhang *et al.* \(2008\)](#); [Xu \(2009\)](#); [Sheikh and Shah \(2005\)](#); [Zhong and Sclaroff \(2003\)](#); [Yang and Chen \(2012\)](#).

2.1.1.2 Classification

After background subtraction, a number of false positives such as shadows are often recognized as the foreground. Some standard filtering techniques based on connected component analysis, size, median filter, morphology, and proximity can be used to improve the performance ([Cucchiara *et al.*, 2003](#); [Elgammal *et al.*, 2000](#); [Guha *et al.*, 2005](#); [McKenna *et al.*, 2000](#); [Yang *et al.*, 2005b](#); [Zhao and Nevatia, 2004a](#)). Methods to directly identify the incorrect pixels are also developed. In these methods, classifiers are used to divide the pixels into a number of sub-classes such as unchanged background, shadows, highlights, moving objects, shadow casted from moving object etc. ([Chen *et al.*, 2005](#); [Cucchiara *et al.*, 2003](#); [Horprasert *et al.*, 1999](#)). Classifiers have been built on the basis of gradients ([McKenna *et al.*, 2000](#)), flow information ([Cucchiara *et al.*, 2003](#); [Doyle *et al.*, 2013](#)), and hysteresis thresholding ([Eng *et al.*, 2003](#)).

Unlike algorithms using pixels (or small image patches) to determine the presence of shadows and then extending these pixel-wise solutions to neighbouring areas, [Amato *et al.* \(2011\)](#) partitioned each object area into a set of segments using a simple graph-based method. The segments are then classified as foreground or shadow by analyzing the intrinsic parameters of the segments.

2.1.1.3 Background Updating

Background updating is required in the cases of outdoor scenes, since the background will change with time. For the slow changes in the scenes, background can be updated recursively by combining the current pixel value with a specified weight into a model ([Cucchiara *et al.*, 2003](#); [Elgammal *et al.*, 2000](#); [McKenna *et al.*, 2000](#); [Stauffer and Grimson, 1999](#)). Alternatively, the overall average change in the scene can be measured compared to the expected background. The information is then used to update the background model ([Fehl *et al.*, 2006](#); [Yang *et al.*, 2005b](#)). In a system without real-time requirements, the pixel values from both the past and the future can be employed to update the background ([Figueroa *et al.*, 2006](#)).

To handle rapid changes in a scene, a new mode can be added into the model. For example, in a MoG model, a new mode is captured by a new Gaussian distribution whenever a non-background pixel is detected. This distribution is given more weighting if more pixels support it. A similar approach is proposed in [Fehl *et al.* \(2006\)](#); [Kim *et al.* \(2005\)](#); [Ilyas *et al.* \(2009\)](#); [Geng and Xiao \(2011\)](#); [Li *et al.* \(2010b\)](#) where each pixel in the background model is represented by a number of codewords, the collection of which is known as the codebook. During run-time a new codeword is created for each foreground pixel. For a certain codeword, if no pixels are required to be assigned to it in a number of frames, the codeword will be removed.

2.1.1.4 Background Initialization

A background model is learned in an initialization phase. In earlier approaches, it is assumed that there is no moving object in several consecutive frames initially, thus the model can be learned from these frames. However, this assumption is not always applicable in real scenarios thus methods of initialization under the presence of moving objects are developed.

To some extent, the MoG representation allows the presence of moving objects during initialization since each moving object can be represented by its own distribution with a low weighting. However, this distribution is likely to generate false positive during the process of classification. Another approach is to collect the pixels which are recognized as genuine background pixels under certain conditions. The background model is initialized using only these collected pixels ([Eng *et al.*, 2003](#); [Gloyer *et al.*, 1995](#); [Haritaoglu *et al.*, 2000](#)).

There are some alternative approaches. The pixels from a number of consecutive frames in the initialization phase are divided into temporal subintervals with similar values. The optimal subinterval where pixels are most likely to belong to the background are identified as the subinterval with the minimum average motion (measured by optical flow) ([Gutchess *et al.*, 2001](#)) or the subinterval with the maximum ratio between the number of samples in the subinterval and their variance ([Wang and Suter, 2005, 2006](#)). In codeword method, a temporal filter is used to eliminate any codeword that has not recurred for a long period of time.

2.1.2 Motion-based Approaches

Motion-based human detection and tracking is based on the idea that differences in consecutive frames are resulting from moving humans. Thus, the human can be found by detecting the motion. The motion in a sequence can be measured based on either the flow or image differencing. In [Sidenbladh \(2004\)](#), the features for a number of image windows containing a walking human are extracted based on the optical flow. A trained support vector machine (SVM) is used to detect walking humans in video. Since optical flow is noisy, higher level entities can be used to measure image flow. For example, KLT-features ([Gonzalez *et al.*, 2003](#)) or displacements of pixel-blocks ([Sangi *et al.*, 2001](#)) are used to extract flow vectors. In [Bradski and Davis \(2000\)](#); [Yin and Collins \(2006\)](#), flow vectors are described by the gradients in motion history images (MHI) ([Davis and Bobick, 1997](#)). In [Ming-yu and Hauptmann \(2009\)](#), the MoSIFT descriptor is proposed to detect spatially distinctive points of interest with substantial motions. In this descriptor, the well-known SIFT algorithm is applied to find visually distinctive components in the spatial domain and detect spatio-temporal points of interest with (temporal) motion constraints. The motion constraint consists of a ‘sufficient’ amount of optical flow around the distinctive points. The motion information in the MoSIFT descriptor is used in [Garcia-Martin *et al.* \(2011\)](#) to characterize the movements and build a motion model to track human.

Image differencing adapts quickly to changes in the scene, but pixels from a human area that have not moved or are similar to their neighbours are difficult to detect. Therefore, three consecutive images ([Haritaoglu *et al.*, 2000](#)) are usually used for image differencing. In [Viola *et al.* \(2003\)](#), a sophisticated image differencing method is introduced based on the principle of their novel face detector ([Viola and Jones, 2001](#)) where a number of simple features are organized in a cascade of progressively more advanced classifiers. A rectangle of pixels is compared between the current image and the previous image. The rectangle from the pervious image is shifted up, down, left, and right in the current image. If the energy is lower in the output in a certain direction, the probability that the human has actually moved (shifted) in this direction is higher. Although the rectangle features of pixels have been successfully applied to face detection ([Viola and Jones, 2001](#)), the results for human detection are not satisfactory ([Dalal and Triggs, 2005](#); [Viola *et al.*, 2003](#); [Zhu *et al.*, 2006](#)). To solve the problem, gradient-based features ([Chen and Chen, 2008](#)) are introduced to increase the discriminating power of the features for human detection.

2.1.3 Appearance-based Approaches

Appearance-based segmentation is built upon two observations: 1) humans have different appearances to the background; and 2) different individuals have different appearances. After a human appearance is built, the segmentation is usually implemented by extracting the appearance of the segmented foreground objects in the current image and then comparing them with the known appearance model, or by directly detecting the pixels in the current image belonging to the known appearance model. These methods can be divided into two categories: independent on temporal context (temporal context-free) where a generic appearance model is adapting to any sequence; or dependent on temporal context where an appearance model is learned or updated in the current sequence.

2.1.3.1 Temporal Context-Free

Temporal context-free methods can be used for detecting humans in a still image (Mohan *et al.*, 2001), for determining humans entering a scene (Okuma *et al.*, 2004), or for indexing images in databases (Burak Ozer and Wolf, 2002). In Okuma *et al.* (2004), a massive amount of training data is used to train an AdaBoost-based classifier. In Utsumi and Tetsutani (2002), the image is divided into a number of blocks and each block is represented by the mean and covariance matrix of the intensities. A distance matrix is constructed to represent the generalized Mahalanobis distance between two blocks. The detection is implemented based on the fact that for non-human images the distances between blocks in the proximity will be larger than those for images containing a human. Covariance features were introduced in Tuzel *et al.* (2006) for matching and texture classification problems, and later extended by Porikli *et al.* (2006) and Tuzel *et al.* (2007) for tracking. A region can be represented by the covariance matrix of image features, such as spatial location, intensity, higher order derivatives, etc. In Alahi *et al.* (2010), any object of interest is described by a cascade of grids of region descriptors. The region descriptors could be the covariance matrix of various features, the histogram of colours, the histogram of oriented gradients, the scale invariant feature transform (SIFT), the speeded-up robust features (SURF) descriptors, or the colour of the interest points. According to different region descriptors, the corresponding effective distance measurement is adopted.

2.1.3.2 Temporal Context

The methods based on temporal context refer to those methods where an appearance model is learned or updated in the current sequence. These methods either operate at pixel level or region level. At pixel level, after an appearance model is built, the likelihood of each pixel conforming to the appearance model is calculated to detect the foreground pixels. At region level, after an appearance model is built, a region in an image is checked by the appearance model to obtain the probability of the region corresponding to the particular appearance model.

Colour-based appearance models have received wide attention. Usually, the colour of a human is represented by either a colour histogram (Comaniciu *et al.*, 2003; Hu *et al.*, 2004; McKenna *et al.*, 2000; Okuma *et al.*, 2004; Xu and Puig, 2005; Zhao and Nevatia, 2004b; Zhao *et al.*, 2008) or a MoG (Kang *et al.*, 2005; Taylor, 2000; Yang *et al.*, 2005a; Jin and Qian, 2009). The comparison between colour histograms normally uses the Bhattacharyya distance and the effect of comparison can be improved by weighting pixels close to the centre of the human higher than those close to the border (Comaniciu *et al.*, 2003; Zhao and Nevatia, 2004b). In Zhao and Nevatia (2004b); Zhao *et al.* (2008), the similarity is combined with the dissimilarity of the background colour histogram. Mahalanobis distance can be used to compare the MoG representations. The efficient evaluation can be obtained by using only one Gaussian (Kang *et al.*, 2005) or by assuming independence between colour channels (Cucchiara *et al.*, 2004). Alternatively, only the mean is used to compare the MoG representations by Yang *et al.* (2005a).

The major challenge of appearance-based tracking is the frequent presence of visual occlusions. Occlusions make the current observation totally or partially unavailable in some time intervals. Many works addressed the occlusion problems. Generally, the approaches of appearance-based tracking attempt only to detect occlusions. The appearance model is not updated when the occlusions happen. Occlusion likelihood is measured by the ratio between the numbers of the observable points and the points of the appearance models. In Zhao and Nevatia (2004a); McKenna *et al.* (2000), this ratio determines whether the object is partially or totally occluded. In Vezzani *et al.* (2011), the occlusions are not only detected but also classified for appearance model updating.

Using only one colour appearance model to represent the whole human body is generally too coarse. To generate a more sophisticated representation for the whole human, spatial information is introduced by dividing the human into a number of sub-regions and representing each sub-region with a colour histogram or a MoG (Mittal and Davis, 2003;

Okuma *et al.*, 2004; Taylor, 2000; Yang *et al.*, 2005a). In Hu *et al.* (2004), an approach is proposed to have three sub-regions representing the head, torso, and legs respectively. In Park and Aggarwal (2006), a human is modelled as a number of blobs where each blob is a group of connected pixels having a similar colour.

2.1.4 Shape-based Approaches

Due to the particular structure of the human body, the shape of a human can be differentiated from the shape of other objects in a scene. The shape of human can therefore become a powerful cue to detect human in an image. In contrast to the appearance-based models, the shape-based models of different individuals are similar. Hence, the shape-based methods are applicable in tracking problems only involving simple correspondences. As with the appearance-based approaches, shape-based approaches are also divided into two categories: temporal context-free and temporal context.

2.1.4.1 Temporal Context-Free

In Leibe *et al.* (2005), the outlines of walking humans are learned and stored in a number of templates. These templates are matched with the edge feature of an input image in different scales based on the Chamfer matching. In Wu and Yu (2006), a human shape model based on human edges is learned and represented as a Boltzmann distribution in a Markov Field. Different locations, scales, and rotations are searched by a detector implemented using a Particle Filter. In Dalal and Triggs (2005), humans are detected by an SVM in a window of pixels. The features in the input image are extracted by a spatially arranged set of HoG (histogram of oriented gradients) descriptors. The input image regions are divided into a number of cells. Each cell is represented by a 1D histogram of gradient directions over the pixels in the cell. This work is extended in Dalal *et al.* (2006) by including motion histograms. It allows human detection even when the camera and/or background is moving. HoGs are combined with Shape Contexts (Belongie *et al.*, 2001) and SIFT (i.e., scale invariant feature transformation) (Lowe, 2004). In Zhao and Davis (2005), a hierarchy of silhouette templates are learned for the upper body. Sitting humans in a frame are detected using the templates storing the outline of certain silhouettes. It is achieved by the assistance of Chamfer matching over different scales and a colour-based detector which is updated iteratively. Shapelet features (a set of informative mid-level features) are proposed by Sabzmeydani and Mori (2007) to discriminate between pedestrian and non-pedestrian classes. A shapelet feature is constructed by selecting a

subset of its low-level features using AdaBoost. These low-level features are generated by extracting the gradient responses of each image in different directions and computing the local average of these responses around each pixel. Adaptive contour feature (ACF) is proposed in [Gao *et al.* \(2009\)](#) for human detection and segmentation. This feature is composed by a chain of a number of granules in oriented granular space (OGS) that is learned via the AdaBoost algorithm. To automatically mine object contour features and feature co-occurrences, three operations are defined on OGS. A weak classifier is defined for human detection or segmentation by generating an ACF.

2.1.4.2 Temporal Context

When the temporal context is taken into account, the shape-based methods can be employed to track humans over time. In situations of temporal smoothness, the shape in the previous frame can be used to detect a human in the current frame. In [Haritaoglu *et al.* \(2000\)](#), a binary edge correlation is performed between the outlines of the silhouettes in the last frame and the surroundings in the current frame. A point distribution model (PDM) is used by [Davis *et al.* \(2000\)](#) to represent the outline of the human. The most likely configurations of humans in the last frame are used to predict the locations of humans in the current frame using a particle filter. Predictions are evaluated by comparing the edge features of the outline with those in the image. Similarly, the active shape model is proposed in [Koschan *et al.* \(2003\)](#). In [Atsushi *et al.* \(2002\)](#), the pose of the human in the previous frame is modelled by ellipses. Based on this ellipse model, nine possible poses of the human are predicted for this human in the current frame and the final pose in the current frame is determined by correlating these poses with the silhouettes in the current frame. In [Krüger *et al.* \(2005\)](#), a hierarchy of silhouettes of walking persons is learned and correlated with the extracted silhouette. A Bayesian tracking framework is used to estimate the translation, scale, and type of the silhouette. In [Munder *et al.* \(2008\)](#), a human is represented by combining a generative shape model and a discriminative texture classifier, both consisting of a mixture of pose-specific submodels. A set of linear subspace models that is an extension of the point distribution models is used to represent the shape features where the shape transitions are modelled by a first-order Markov process. Texture is represented by a manifold that is implicitly delimited by a set of pattern classifiers where texture transition is modelled by a random walk. Object detection and tracking is achieved by employing a Bayesian framework based on particle filtering. In [Li *et al.* \(2010a\)](#), HoG (histogram of oriented gradients) is used to extract human feature in images and the prediction and estimation from the Kalman filter are introduced to assist human detecting and tracking. For cases of partial occlusion the shape-based method would easily fail since the global shape information becomes unavailable. A whole human is hence

proposed to be divided into a few parts to handle the partial occlusion problem. In [Wu and Nevatia \(2005\)](#), four different body parts including the full-body, head-shoulder, torso, and legs are separately detected. For each body part, a detector is trained based on a boosting classifier combined with the edgelets (small connected chains of edge pixels). In the cases of people grouping together, the occlusions often happen and most of the time the only reliably shape information is the head or head-shoulder profile.

2.1.5 Section Summary

In this section, some typical approaches for human detection and tracking are reviewed, which include background subtraction, motion-based approaches, appearance-based approaches and shape-based approaches. These approaches are usually used as the pre-processing step in the system of human pose estimation to estimate a coarse position of tracked human and thus effectively reducing the search space for human pose estimation. In the following sections, the approaches for human pose estimation will be reviewed.

2.2 Pose Estimation: Model-free Approaches

This category encompasses approaches where a direct relation between image observations and human postures is built, rather than matching a pre-defined human model to image observations. Two types of approaches have been proposed which fall into this category: learning-based approaches and example-based approaches. In learning-based approaches, a mapping from the image space to the pose space is learned from training data. In example-based approaches, a set of exemplars for human poses is selected. Matching indexes between the pose descriptions and the image observations are built and stored in a database. For a given input image, a search is performed based on the corresponding index and the candidate poses are interpolated to form the final pose estimation.

2.2.1 Learning-based Approaches

[Grauman *et al.* \(2003\)](#) proposed a learning-based approach to estimate a 3D human pose from multi-view silhouettes using a probabilistic ‘shape+structure’ model. In this approach, a human silhouette in a certain view is represented by a shape vector composed of a set of sampled points on the closed contour, which is a global descriptor. A feature

vector is formed by concatenating the shape vectors for each view and the vectors for 3D joint locations. A large set of training samples is obtained from multi-view synthetic data where each training sample is composed of a feature vector specifying multi view silhouettes and 3D joint positions. A distribution over the observation space for the true underlying contours together with their associated 3D joint locations is approximated by a mixture of Gaussian models which learns from the training samples via the EM algorithm (Expectation-maximization algorithm). Finally, given the observable contour data, the associated 3D joint locations are estimated by finding the MAP estimation (Maximum A Posteriori estimation) based on the learned mixture of Gaussian models.

The shortcoming of this approach is that any local distortion in the silhouette shape can pollute the global descriptor. Unfortunately, the silhouettes extracted from real image data tend to have distortions in their local form due to factors such as shadow attachment and poor background segmentation. Therefore, to ensure the silhouettes used in the training samples to be clear and clean, they are extracted from synthetic image data rather than from real image data.

Agarwal and Triggs (2006) use a non-linear regression to model the relations between the histograms of shape contexts and 3D poses. In order to improve the resistance to local distortion of silhouette, unlike Grauman *et al.* (2003) where a silhouette shape is used as a global descriptor, local descriptors are introduced to represent a silhouette shape in this approach. A set of local descriptors (shape contexts) is first computed at regularly spaced points on the edge of the silhouette. Shape contexts are used to encode the local silhouette shape at a range of scales, over the regions of diameter similar to the length of a limb. The scale of the shape contexts is determined by a function of the overall silhouette size, making the representation invariant to the overall scale of a silhouette. The shape contexts are composed of $12 \text{ angular} \times 5 \text{ radial}$ bins, resulting in 60-dimensional histograms. The local silhouette shape is thus encoded as a 60D distribution in the shape context space. The representation for the whole silhouette shape is composed of the representation for each local silhouette shape, giving rise to a high dimensional distribution in the shape context space. Matching silhouettes are therefore transformed to matching distributions in shape context space. To implement the matching, the distribution of each silhouette is reduced to a 100D histogram by vector quantizing the shape context space. The 3D body pose is represented by joint angles. Finally, the non-linear regression is used to build the relation between the histograms of shape contexts and the 3D poses.

The two aforementioned approaches can only recover a relatively limited set of predefined activities, e.g., running or walking rather than arbitrary poses due to the multi-modality of the mapping between the observation space and pose space. Some approaches address the

problem by learning a complex appearance-to-pose mapping for arbitrary motions using probabilistic regression. For instance, [Sminchisescu *et al.* \(2005\)](#) proposed a discriminative conditional model representing multi-modal mappings with a mixture of experts (e.g., Gaussian kernel regressors) and [Urtasun and Darrell \(2008\)](#) handle multi-modality by taking advantage of Gaussian Process models.

2.2.2 Example-based Approaches

Example-based approaches first build a set of example poses with their corresponding visual appearance. Pose recovery is then achieved by simply selecting the pose that corresponds to the most visually similar example. Since the examples usually cover the pose space very sparsely, the pose estimate is often obtained through interpolation of multiple close examples.

[Mori and Malik \(2006\)](#) proposed an approach to characterize a shape by a set of sample points from the external and internal contours of an object, found using an edge detector. In an inference step, the shapes of the stored exemplars are deformed to match the image observation. During this deformation, the locations of the manually-labelled 2D body joints also change. The most likely 2D joint estimation is found by enforcing the 2D image distance consistency between body parts.

2.2.3 Section Summary

Model-free approaches are computationally efficient, and could potentially apply in real-time, once the relations between the image observations and the human postures have been built. However, the estimation accuracy of model-free approaches is not as good as that of model-based approaches. Human pose estimation from images is challenging due to the large variations in human body dimensions, camera viewpoint, type of motion and numerous environmental settings such as lighting. For model-based approaches, these variations can be parameterized and accurately represented by a group of flexible models. It means that the accurate human pose is possible to be recovered in the estimation stage. As for the model-free approaches, all these variations can only be roughly represented by a limited set of training data which implies that model-free approaches are unable to accurately estimate human pose, especially when there are large variations in the input image.

2.3 Model-based Pose Estimation: Top-down Approaches

Top-down approaches can also be called the model-based analysis-by-synthesis approaches. This type of approach first builds a human model consisting of an explicit geometric representation of human shape and its kinematic structure. Human pose is then estimated by optimizing the similarity between the model projection and the observed images.

In top-down estimation approaches, the problem of human pose estimation over time is generally formulated as the computation of the posterior probability distribution $p(X_t|I_{1:t})$ over a parameter vector (X_t) of the human model, given a sequence of images ($I_{1:t}$). This can be expressed as a marginalization of the joint posterior over all states ($X_{1:t}$) up to time t given all images ($I_{1:t}$) up to time t :

$$p(X_t|I_{1:t}) = \int p(X_{1:t}|I_{1:t})dX_{1:t-1}. \quad (2.1)$$

Using Bayes' rule and the Markov assumptions, it can be shown that the dependence on states at times before time $t - 1$ can be removed, to give

$$p(X_t|I_{1:t}) \propto p(I_t|X_t) \int [p(X_t|X_{t-1})P(X_{t-1}|I_{1:t-1})]dX_{t-1}. \quad (2.2)$$

Here, $p(I_t|X_t)$, which we refer to as the observation likelihood, is the probability of the image being observed at time t , given the human configuration parameter states (X_t) at time t . The integral in Equation (2.2) is referred to as a prediction, as it is equivalent to the probability over states X_t at time t given the image measurement history $I_{1:t-1}$; i.e., $p(X_t|I_{1:t-1})$. It is useful to understand the integrand as the product of two terms: the posterior probability distribution over states at the previous time step, $p(X_{t-1}|I_{1:t-1})$, and the dynamical process, $p(X_t|X_{t-1})$, that propagates this distribution over states from $t - 1$ to t .

To estimate human pose under this Bayesian formulation, a human model must first be defined to determine the parameter vector (X_t). There are various definitions for the human model according to different practical requirements. Some typical definitions of human model will be reviewed in Section 2.3.1. Computation of the posterior distribution is difficult due to the nonlinearity of the likelihood function $p(I_t|X_t)$ over human model parameters. Although an analytic expression for the likelihood function over the parameters of the entire state space cannot be derived, the likelihood of observing the image can be evaluated given a particular state (X_t). We will review a few typical approximation approaches of likelihood function in Section 2.3.2. In practice, the dynamic process, which

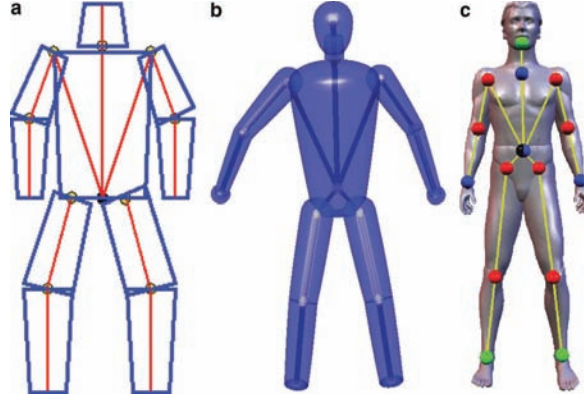


Figure 2.1: Human shape model with kinematical model. (a) 2D model (reprinted from [Huang and Huang \(2002\)](#)); (b) 3D volumetric model consisting of superquadrics (reprinted from [Kehl and Gool \(2006\)](#)); (c) 3D surface model (reprinted from [Carranza *et al.* \(2003\)](#)).

is called as motion prior, is used to capture the characteristics of human activities. Several definitions of motion prior will be reviewed in Section 2.3.3. Finally, many approaches have been proposed to compute the posterior probability distribution, which will be reviewed in Section 2.3.4.

2.3.1 Human Model

Human models generally consist of two components: a geometric representation of the human shape and its kinematic structure. In most models, the skeletal structure is described as a kinematic tree, which is made of a collection of segments that are linked by joints. Every joint can be considered to have a number of degrees of freedom (DOFs). All DOFs in the human model together form the pose representation. The number of DOFs for the same joint in different models could be different, hence the number of the total DOFs for all body parts can be varied in different human models, e.g., 23 DOFs in [Wachter and Nagel \(1997\)](#) and 25 DOFs in [Sidenbladh *et al.* \(2000\)](#).

Beside the skeletal structure, a geometric representation of a human shape also needs to be defined. As shown in Figure 2.1(a), the shapes of body parts in a 2D human model are usually described as rectangular or trapezoid-shaped patches. In 3D human models, the shapes of body parts are usually modelled by volumetric models such as spheres ([O’Rourke and Badler, 1980](#)), cylinders ([Hogg, 1983](#); [Rohr, 1994](#); [Sidenbladh *et al.*, 2000](#)) or tapered super-quadrics ([Kehl and Gool, 2006](#)) (See Figure 2.1(b)). Instead of modelling each body part as a separate rigid shape, the surface of the entire human body can also be modelled by

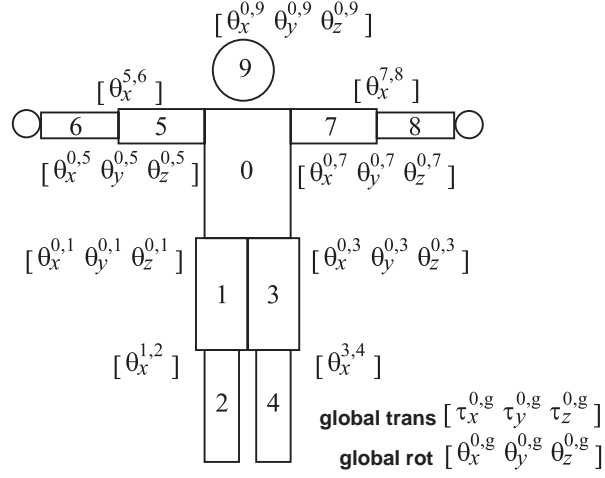


Figure 2.2: 3D kinematical model of [Sidenbladh et al. \(2000\)](#)

surface-based models (See Figure 2.1(c)) where polygonal meshes are used to approximate the surface of the entire human body. The deformations of these polygons are controlled by the underlying kinematic model during motion. It is expected that the more complex the models, the better the tracking results. However, complex human model require more parameters to be estimated, which will lead to more expensive computation during the matching process, and very often more uncertainties.

A typical example of a human model definition is described as follows. [Sidenbladh et al. \(2000\)](#) modelled the human body as a configuration of nine cylinders and three spheres. Any two connected body parts are connected by a joint. Normally, beside a global co-ordinate system, each body part has a local coordinate system with its origin located at the corresponding joint position. When defining a kinematic model, the accompanying coordinate systems and the transformations between the local coordinate systems will be defined. Transformations between coordinate systems are expressed using homogeneous transformation matrices:

$$T = \begin{bmatrix} R_z R_y R_x & t \\ 0 & 1 \end{bmatrix} \quad (2.3)$$

To order the transformations between the coordinate systems of different limbs, a kinematic tree as shown in Figure 2.2 (The body parts are numbered for ease of identification) is defined where the torso is specified as the root node. For example, a point P on limb 1 (the left thigh) with a coordinate P_1 in the local coordinate system of limb 1 can be transformed

to the corresponding coordinate P_g in the global coordinate system by $P_g = T_{0,g}T_{1,0}P_1$. The transformation from the left thigh coordinate system to the torso coordinate system are represented by $T_{1,0}$, while the transformation from the torso coordinate system to the global coordinate system is represented by $T_{0,g}$. The entire pose of the human body can be represented by 25 parameters (25 DOFs), i.e., the angles at the shoulders, elbow, hips and knees, and the position and orientation of the torso in the scene.

2.3.2 Likelihood Function

The computation of the posterior distribution is difficult due to the nonlinearity of the likelihood function $p(I_t|X_t)$ over human model parameters. Specifically, an analytic expression for the likelihood function over the parameters of the entire state space cannot be derived. Instead, the likelihood of observing the image given a particular state (X_t) is evaluated. In the existing approaches, the likelihood function is usually built based on visual hull (Caillette *et al.*, 2005; Mikic *et al.*, 2001; Chari *et al.*, 2012) or foreground and edge images (Deutscher and Reid, 2005; Sminchisescu and Jepson, 2004; Bae *et al.*, 2013). The visual hull methods tend to be faster but inherit the visual hull’s sensitivity to segmentation errors. In contrast, the foreground/edge images method is more robust against segmentation errors. Thus in recent years the likelihood functions in most approaches are built based on foreground/edge images. The most typical likelihood function based on foreground/edge images is proposed by Deutscher and Reid (2005), with some minor modifications in Balan *et al.* (2005) and Peursum *et al.* (2007).

In Deutscher and Reid (2005), the likelihood function $p(I_t|X_t)$ over human model parameters is approximated by a weighting function which is constructed based on two image features: edges and foreground silhouette. For the edge feature, a gradient-based detector is employed to detect the edges where a threshold is set to eliminate spurious edges. A pixel map (as shown in Figure 2.3 (a)) is produced where each pixel is assigned a value to specify its proximity to an edge. The weighting function $W^e(X_t, I_t)$ based on edge feature is defined as

$$W^e(X_t, I_t) = \frac{1}{N} \sum_{i=1}^N (1 - p_i^e(X_t, I_t))^2, \quad (2.4)$$

where X_t is the configuration vector of the human model at time t and I_t is the image from which the pixel map is derived. $p_i^e(X_t, I_t)$ are the values of the edge pixel map at the N sampling points taken along the human silhouette. For the foreground silhouette, the silhouette feature is extracted by background subtractions which separates the subject from the background. Once again a pixel map (as shown in Figure 2.3 (b)) is constructed where the foreground pixel is set to 1 and background pixel is set to 0. The weighting

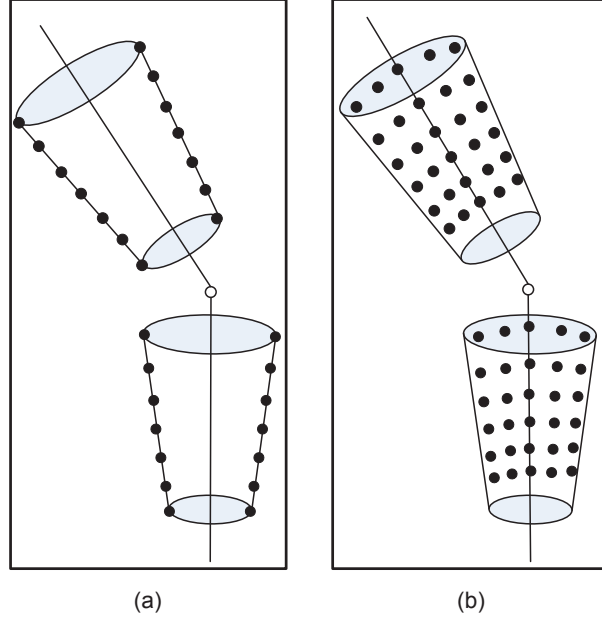


Figure 2.3: Configurations of the pixel map sampling points $p_i(X_t, I_t)$ for the edge based measurements (a) and the foreground segmentation measurements (b).

function $W^r(X_t, I_t)$ based on the silhouette feature is defined as

$$W^r(X_t, I_t) = \frac{1}{N} \sum_{i=1}^N (1 - p_i^r(X_t, I_t))^2, \quad (2.5)$$

where $p_i^r(X_t, I_t)$ are the values of the foreground pixel map at the N sampling points taken from the interior of the truncated cones. The weighting function combining both the edge and silhouette features is given as

$$w(X_t, I_t) = e^{-(W^e(X_t, I_t) + W^r(X_t, I_t))} \quad (2.6)$$

In practice, the inaccurate evaluations from the observation likelihood are inevitable due to cluttered environments, occlusions and low-level algorithm failures. Unfortunately, when inaccurate evaluations occur, the correct posture becomes less likely than other seemingly plausible postures and tracking failure becomes possible. In the approaches built based on multiple views, the observation likelihoods from multi views can be combined to improve the accuracy of the observation likelihood thus reducing the probability of the occurrence of tracking failure to some extent. In approaches built based on monocular views where no extra observation likelihoods can be used, the main idea to avoid tracking failure is to narrow the size of the posture space which can eliminate a part of the seemingly plausible postures. The size of the posture space is decided by a motion model which will be reviewed in the next section. It is notable that some existing approaches designed for

arbitrary human activities are commonly built based on multiple views since their posture space cannot be narrowed down in order to cover more human activities (Deutscher *et al.*, 2000; Deutscher and Reid, 2005).

2.3.3 Dynamic Model

In the top-down approaches, the characteristics of human activity is usually captured by a motion prior $P(X_t|X_{t-1})$ which is used to predict the current posture based on the previous posture. The posture space for the current pose will be decided by the motion prior and the previous posture. Different motion priors will correspond to posture spaces with different sizes. For instance, if a motion prior is trained to capture the characteristics of arbitrary human activities, the size of its corresponding posture space will be larger than the size of a posture space corresponding to a motion prior which captures only the characteristics of a specific human activity such as walking and running. As discussed before, inaccurate evaluations from the observation likelihood are inevitable, and this can lead to tracking failure. Although it is impossible to completely prevent tracking failure, the probability of its occurrence can be reduced by narrowing down the size of posture space via motion models.

Deutscher and Reid (2005) uses a generic motion prior which searches for the current posture located in the neighbouring space of the previous posture. Although this motion prior can be applied in arbitrary human activity, it is unable to effectively restrict the size of the posture space thus easily resulting in false estimation. This type of motion prior is normally applied to approaches based on multiple views, which can provide a more robust observation likelihood. It is not suitable for applications based on monocular view.

In comparison with Deutscher and Reid (2005), Balan *et al.* (2005) described a motion prior from a training set focusing on the walking motions, which will restrict the practical size of posture space. Peursum *et al.* (2010) built a more sophisticated motion prior by modelling the motions with a variant of the hierarchical hidden Markov model (HMM), which further narrows the size of the posture space.

2.3.4 Posterior Estimation

A wide range of estimation techniques have been proposed for model-based pose estimation such as Kalman filter (Bregler and Malik, 1998; Ligorio and Sabatini, 2013), the condensa-

tion algorithm (Isard and Blake, 1998), and dynamic Bayesian network (BN) (Sidenbladh *et al.*, 2000). Conventionally, Kalman filtering and its variations are known for its efficiency and capability of accurate posterior estimation. However, it is an estimation method based on the Gaussian distribution, and is therefore restricted to situations where the probability distribution of the state parameters is uni-modal. In order to cope with clutter situations in which modelling parameters of probability density functions (PDFs) are usually multi-modal and non-Gaussian, stochastic sampling strategies are designed to represent simultaneous alternative hypotheses. Among the state-of-the-art in stochastic sampling approaches, the condensation algorithm (which is also called particle filter) is the dominant method (Sidenbladh *et al.*, 2000). It is based upon sampling the posterior distribution estimated in the previous frame, and propagates these samples iteratively to successive images.

The posterior distribution $p(X_t|\bar{I}_t)$ is represented by a set of weighted particles $\{(s_t^{(0)}, \pi_t^{(0)}) \dots (s_t^{(N)}, \pi_t^{(N)})\}$ where the weights $\pi_t^{(n)} \propto p(I_t|X_t = s_t^{(n)})$ are normalized so that $\sum_N \pi_t^{(n)} = 1$. The state \mathcal{X}_t at each time step t can be estimated by

$$\mathcal{X}_t = \mathcal{E}_t[X_t] = \sum_{n=1}^N \pi_t^{(n)} s_t^{(n)} \quad (2.7)$$

or the mode

$$\mathcal{X}_t = \mathcal{M}_t[X_t] = s_t^{(j)}, \pi_t^{(j)} = \max(\pi_t^{(n)}) \quad (2.8)$$

of the posterior distribution $p(X_t|\bar{I}_t)$.

Particle filtering works well because it can model uncertainty under non-linear and non-Gaussian conditions. Less likely model configurations will not be discarded immediately but given a chance to prove themselves later on, resulting in more robust tracking. However a price needs to be paid for these attributes in computational cost. The most expensive operation in the standard condensation algorithm is the evaluation of the likelihood function $p(I_t|X = s_t^{(n)})$ and this has to be done once at every time step for every particle. To maintain a fair representation of $p(X|I_{1:t})$ a certain number of particles are required, and this number grows exponentially with the dimensionality of the model's configuration space. Therefore, the fundamental difficulty with the particle filter techniques to human pose estimation is the dimensionality of the state space.

Deutscher *et al.* (2000) introduced the annealed particle filter which combines a deterministic annealing approach with stochastic sampling to reduce the number of samples required. At each time step the particle set is refined through a series of annealing cycles with decreasing temperature to approximate the local maxima in the fitness function.

Whilst effective, it is susceptible to being caught in poor modes due to its concentration of particles within a found mode.

2.3.5 Section Summary

Along with the introduction of stochastic sampling and search techniques, top-down approaches have achieved human pose estimation without constraint on human activity. For instance, as discussed in Section 2.3.3, Deutscher *et al.* (2000) used a generic motion prior to represent any type of human motion. Their approaches can be applied to estimate human pose from any activity. However it requires the observations from multiple views. As for other approaches based on monocular view, such as the approach proposed by Peursum *et al.* (2010), they are able to accurately estimate human pose over time, but is limited to specified human motions, such as walking and running. According to previous analysis, we can see that applying top-down approaches to estimate human poses over time requires either the accurate observation likelihood from multiple views or an effective motion prior trained for a specific human activity.

2.4 Model-based Pose Estimation: Bottom-up Approaches

Bottom-up human tracking approaches (Felzenszwalb and Huttenlocher, 2005; Crandall *et al.*, 2005; Crandall and Huttenlocher, 2006; Sigal *et al.*, 2003; Lan and Huttenlocher, 2005; Ren *et al.*, 2005; Shotton *et al.*, 2013b; Eichner *et al.*, 2012) are essentially designed to locate body parts and then assemble them into a human body. Although this type of approach is typically applied for estimating human configuration in a single frame, it can be easily extended for estimating human configuration over time by detecting each frame in a video sequence independently. A human motion model is optional in this kind of approach, hence they can be used in tracking unconstrained movements.

2.4.1 Segmentation-based

In Mori *et al.* (2004), image segmentation is performed first to generate candidate segments. A set of low-level cues including contour cue, shape cue and shading cue is then computed to classify these segments. For the contour cue, the contour of a segment is measured to determine how well-separated it is from the background. For the shape cue,

a rectangle is assumed to capture the basic shape characteristics of half-limbs. According to the size and orientation of a segment, a rectangle template is constructed. The shape cue is then defined as the overlapping area between the segment and this reconstructed rectangle. A sigmoid function is introduced to transform each cue into a probability-like quantity and then combine them linearly. Half-limbs and torso are identified from these segments by the trained body part locators based on the combination of the low-level cues. Normally, a partial configuration can be found, followed by a search for the missing body part(s). To prune the search space, global constraints such as body part proximity, relative widths and lengths of body parts, and symmetry in colour etc. are enforced.

Following Mori *et al.* (2004), a similar approach is proposed by Ren *et al.* (2005), where pairwise edges are searched as segments and integer quadratic programming is used to search human body configuration. No explicit human model is used in this approach. Instead, pairwise constraints between parts are defined to approximate the global configuration consistency.

The two approaches above both share two properties: 1) selecting the candidate segments based on some low-level and simple features; and 2) prematurely filtering the possible body segments at the part detection stage. It facilitates the reduction of computing complexity and search space, while sacrificing the accuracy of human pose estimation. If some crucial segments corresponded to body parts are removed before the search for the human body configuration, the correct human pose is impossible to be estimated. The success of these two approaches is heavily dependent on the success of the low-level segmentation algorithm.

2.4.2 Pictorial Structure-based

Beside approaches based on segmentation, approaches based on a pictorial structure model have become very important for bottom-up approaches. The approaches proposed in this thesis are built upon the pictorial structure model. Generally, a body model can be decomposed into a set of body parts. Their configuration is denoted as $L = \{l_0, l_1, l_2, l_3, \dots, l_n\}$, where n is the number of body parts. Given an image I and an appearance model C , the posterior of a human configuration is modelled as

$$P(L|I, C) \propto P(I|L, C)P(L). \quad (2.9)$$

An estimation of the pose configuration is modelled as an inference problem in the probabilistic model. In this case, body parts can be identified by using an appearance likelihood

model $P(I|L, C)$ to match the image data. The spatial relation between body parts is encoded in the kinematical prior $P(L)$. Note that there is no relationship with the configuration in the previous frame since the estimation is based on a single frame, hence a motion model is not required. The right selection of components for both the appearance and spatial modelling is crucial for the general applicability and overall performance of the pictorial model. There are two important components in the pictorial structure approaches: appearance likelihood model and kinematical prior.

2.4.2.1 Appearance Likelihood Model

Most pose estimation approaches based on the pictorial structure model are typically designed for estimating human pose in a single image. Since different people appear differently in images due to different clothings and body shapes, to estimate different persons' pose by a single approach, a generic appearance model C (Ferrari *et al.*, 2008; Mikolajczyk and Schmid, 2005) is needed to describe the common appearance features of a human figure. Such an appearance model can be built based on generic features such as background subtraction (Felzenszwalb and Huttenlocher, 2005), Gaussian derivative (Ronfard *et al.*, 2002; Jain and Crowley, 2013) and shape context (Andriluka *et al.*, 2009; Shotton *et al.*, 2013a) or other detection means.

Felzenszwalb and Huttenlocher (2005) proposed a pictorial structure to estimate human body pose. In this model, the shapes of most body parts are assumed to be cylindrical. The projection onto an image is thus approximated by rectangles. The width of a rectangle comes from the diameter of the corresponding cylinder which is fixed for a particular person, while the length of the rectangle depends on the length of the cylinder but can vary due to foreshortening effect. The projection of a body part is modelled as a rectangle parameterized by (x, y, s, θ) , where the centre of the rectangle is given in the image coordinate (x, y) , the length is defined by the amount of foreshortening $s \in [0, 1]$, and the orientation is given by θ . To define the appearance likelihood model, the rectangle for each body part is divided into two parts including the central area (Area 1) and the border area (Area 2) as shown in Figure 2.4. The appearance likelihood model $P(I|L, C)$ is built upon a binary image consisting of foreground pixels and the background pixels, which is obtained by background subtraction. Before defining the appearance likelihood $P(I|L, C)$, two parameters q_1 and q_2 are defined. q_1 is the probability that the foreground pixels locating in Area 1 belong to the body part and q_2 is the probability that the foreground pixels locating in Area 2 belong to the body part. The appearance likelihood $P(I|L, C)$

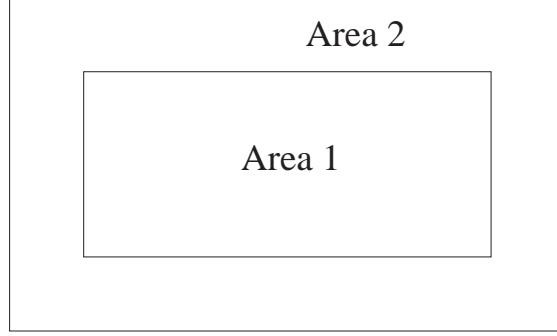


Figure 2.4: A rectangular body part. Area 1 is the central area inside the part, and Area 2 is the border area around Area 1.

is hence modelled as

$$p(I|L, C) = q_1^{count_1} (1 - q_1)^{(area_1 - count_1)} q_2^{count_2} (1 - q_2)^{(area_2 - count_2)} 0.5^{(t - area_1 - area_2)}, \quad (2.10)$$

where $count_1$ is the number of foreground pixels inside Area 1, and $area_1$ is the area of Area 1. $count_2$ and $area_2$ are similar measures corresponding to the border area (Area 2), and t is the total number of pixels in the image. The appearance parameters are $C = (q_1, q_2)$, and these parameters are estimated from the training examples. There are several limitations in this way of building an appearance likelihood model. Most critical is that it can only be applied to binary images which are generated by applying background subtraction to the original images. Thus its performance is directly affected by the performance of the background subtraction, which severely limits its practical applicability.

In [Ronfard et al. \(2002\)](#), part detectors for each body part are learned instead of an appearance likelihood model. Before training, all body parts in the training images are manually labelled. Each designated body part corresponds to a sub-image area from the training images. All sub-images are scaled to be 14×24 pixels for extracting the feature vector. In the 14×24 sub-image, the feature of each pixel is represented by a 6-tuple $\{G, \nabla_x G, \nabla_y G, \nabla_{xx} G, \nabla_{xy} G, \nabla_{yy} G\}$, which is the absolute values of the responses of six Gaussian filters. Therefore, any body part (sub-image) is represented by a 2016 dimensional feature vector. Using the feature vectors in the training set, for each body part, two linear classifiers (part detector) are trained using Support Vector Machine (SVM) and Relevance Vector Machine (RVM) respectively. In their paper, the framework of pictorial structure is proven to work well without background subtraction, but the image feature (Gaussian derivatives) used to train the part detectors are quite simple.

In [Andriluka et al. \(2009\)](#), a more sophisticated part detector is built using a densely

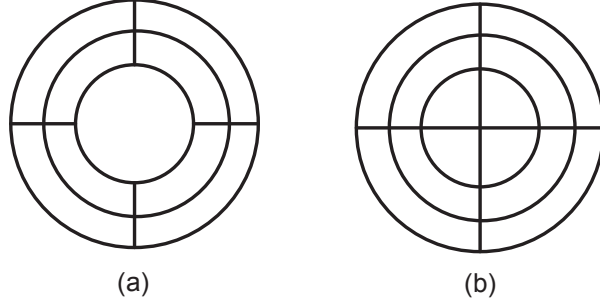


Figure 2.5: (a) The log-polar grid shows 9 location bins used in shape context. (b) 12 location bins used in the variant of shape context.

sampled shape context descriptor and discriminatively trained AdaBoost classifiers. The shape context descriptor was initially proposed in Mikolajczyk and Schmid (2005) and previously used for pedestrian detection (Seemann *et al.*, 2005). The shape context is the 3D histogram of edge point locations and orientations. Edges are extracted by the Canny (1986) detector. In the original shape context descriptor, the location is quantized into 9 bins of a log-polar coordinate system as displayed in Figure 2.5(a) with the radius set to 6, 11 and 15 respectively, and orientation quantized into 4 bins (horizontal, vertical and two diagonals). A 36 dimensional descriptor is therefore obtained. A point contribution to the histogram is weighted with the gradient magnitude. Andriluka *et al.* (2009) uses a variant of shape context descriptor which uses 12 bins for locations as shown in Figure 2.5(b) and 8 bins for gradient orientation resulting in a 96 dimensional descriptor. The feature vector for a body part is built by concatenating all shape context descriptors whose centres fall inside the bounding box for the part, so that some of the feature dimensions can capture the surrounding context. During the detection stage, all possible positions in an image, scales and orientations of the body parts are scanned using a sliding window. An AdaBoost classifier is trained to generate a part detector. In the AdaBoost classifier, a number of decision stumps are used to consider whether one of the log-polar histogram is above a threshold. This generic part detector is effective since the dense appearance representations are computed based on shape context descriptors and AdaBoost is used to train discriminative part classifiers.

2.4.2.2 Kinematic prior

Another important component in the pictorial structure model is the prior $P(L)$, which encodes the probabilistic constraints on part configurations. Such constraints are commonly built based on the kinematic dependencies between parts, hence the prior $P(L)$ is

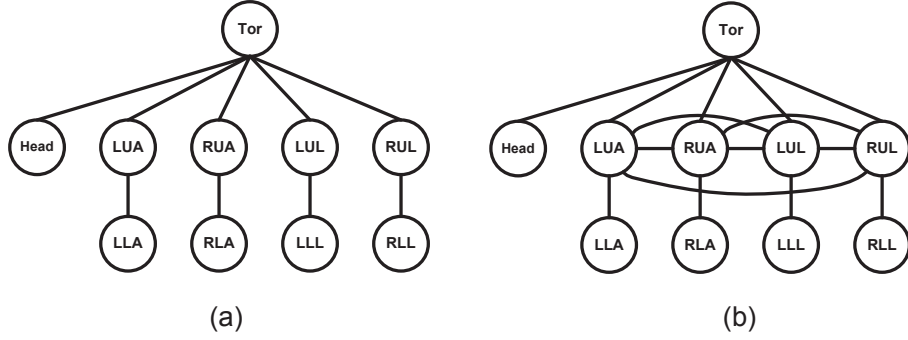


Figure 2.6: (a) Tree model capturing the kinematic model of human body.(b) 12 location bins used in the variant of shape context. (Tor=Torso, LUA=Left-Upper-Arm, RUA=Right-Upper-Arm, LUL=Left-Upper-Leg, RUL=Right-Upper-Leg, LLA=Left-Lower-Arm, RLA=Right-Lower-Arm, LLL=Left-Lower-Leg and RLL=Right-Lower-Leg)

also referred to as kinematic prior. The kinematic structure of a human body is naturally captured by the tree model where the root node represents the torso correlated with the four upper parts, and the four lower parts are connected to their corresponding upper parts, as shown in Figure 2.6 (a). The distribution $P(L)$ for a tree-structured prior can be factorized as

$$P(L) = \prod_{(i,j) \in E} p(l_i, l_j), \quad (2.11)$$

where E denotes the set of all edges in the kinematic tree indexed by the vertices (i,j) that the edge connects. The tree-structured prior model has some limitations. This model can only capture the correlations between body parts connected by joints (such as torso and upper arms etc.), but not the correlations of body parts that are not connected (such as left upper arm and right upper arm etc.). An important property of pictorial structure model based on tree-structured prior is that optimal inference is tractable. An exact model inference can be completed by the max-product algorithm (Felzenszwalb and Huttenlocher, 2005) in linear time.

Besides the tree-structured prior, more complex body models (Crandall *et al.*, 2005; Crandall and Huttenlocher, 2006; Sigal *et al.*, 2003; Lan and Huttenlocher, 2005) have also been studied where the correlations between body parts that are not directly connected by joints are also captured, as shown in Figure 2.6 (b). However, an exact inference algorithm is normally unachievable based on these models. Rather approximate inference approaches such as loopy belief propagation are used which are generally less accurate (Coughlan and Ferreira, 2002).

2.4.3 Section Summary

In this section, some typical bottom-up approaches for human pose estimation are reviewed. In bottom-up approaches, body parts are first detected and then assembled into a human body. Unlike top-down approaches, a motion prior is not necessary for bottom-up approaches. Instead a kinematic prior which encodes the relations between body parts is used to assemble body parts. Unlike the need to learn a motion prior where either multi-view is necessary or human activities needs to be specific, a kinematic prior can be learned independent of human activities, which make it possible to recover arbitrary human poses. In addition, bottom-up approaches are usually designed for estimating human pose in a single arbitrary image which means that the human pose can be estimated under monocular view with unknown camera parameters. Based on this analysis, the bottom-up approach is more suitable to the objective of this thesis than a top-down approach.

In comparison with segmentation-based approaches, pictorial structure-based approaches have more chance to derive an accurate human pose estimation. Unlike segmentation-based approaches which only consider some apparent candidates, the pictorial structure-based approaches do not abandon any possible part candidates before making the final decision. With some reasonable assumptions, the framework of pictorial structure provides an efficient algorithm for achieving a global optimal estimation over all possible locations for each body part and all possible connections between these body parts. Therefore, the framework of pictorial structure is a good choice for our objective in this thesis.

2.5 Pictorial Structure Using a Specific Appearance Model

Two approaches that are very effective in using pictorial structure to detect and track human poses in unconstrained videos are proposed by [Ramanan *et al.* \(2007\)](#) and [Ferrari *et al.* \(2009\)](#). In these two systems, in contrast to some other pictorial structure approaches, a specific appearance model is trained for a particular person rather than using a generic appearance model. Moreover, this specific appearance model is automatically trained based on some generic human detection method. Section 2.5.1 first reviews the approaches proposed by [Ramanan *et al.* \(2007\)](#) and [Ferrari *et al.* \(2009\)](#) and briefly analyzes their limitations in initializing the specific appearance model. In Section 2.5.2, we will analyze why a specific appearance model is more effective than a generic appearance model when tracking a particular person. Finally, Section 2.5.3 analyzes the weakness of the approaches based on a specific appearance model and discuss how to overcome it.

2.5.1 Approaches Review

Ramanan *et al.* (2007) proposed an approach that applies a colour appearance model in the pictorial structure to detect and track the 2D poses of a particular person in consecutive frames with no restriction on human activity. In contrast to a generic appearance model, which can be learned using manually-labelled training samples, a colour (specific) appearance model must be learned based on automatically-labelled training samples since the colour appearance model will change for a different human, whereas a generic appearance model applies to any human. Ramanan *et al.* (2007) proposed to automatically annotate a training set using a ‘stylized pose’ detector which is designed to detect and accurately estimate a lateral walking pose. It is required that a lateral walking pose exists in the video sequence, and the stylized pose detector can be used to accurately detect the lateral walking pose. The stylized pose detector is also built based on the framework of pictorial structure. In this detector, the kinematic prior $P(L)$ is manually set to be uniform within a bounded range consistent with lateral walking motion. For example, the orientation for upper legs is set to be between 45° and 15° with respect to the torso axis for detecting human in a distinctive scissor-leg pattern. This kinematic prior can effectively restrict the size of the posture space thus simplifying the detection and estimation problem. The appearance likelihood model is built upon a chamfer template edge mask (a generic appearance model). Given an image, the chamfer cost of an edge template (which is the appearance likelihood) is the average distance between each edge in the template and the closest edge in the image. To get an accurate estimation, some global constraints are enforced in this detector such as the appearance similarity between symmetric body parts. Assuming that a lateral walking pose appears in at least one frame of a video sequence, the stylized detector can accurately annotate the positions of all body parts of a person in that frame which are used as the training samples to train a discriminative appearance model based on colour. In the event of multiple frames detected with a stylized pose, a clustering step is performed to select the median example as the single training sample. In order to train a colour appearance model, Ramanan *et al.* (2007) trains a quadratic logistic regression classifier where all pixels inside the estimated limb rectangle and all non-person pixels are used respectively as positives and negatives. The appearance model for each limb is a quadratic surface that splits the RGB space into the limb/nonlimbs pixels. The pixels in a frame belonging to each limb can be labelled by the appearance model. The appearance likelihood for a candidate limb is evaluated by summing up the number of misclassified pixels in a local image region specified by the candidate. After colour appearance models for each limb are trained, they are applied in another pictorial structure to estimate human pose in each frame. In this pictorial structure, the spatial kinematics of the human body is modelled with a puppet of rectangles with free-rotating

revolute joints where the distance between the hinge points for the two segments is less than a threshold. Moreover, angular bounds are used to prevent the upper legs from pointing up into the torso. As a final step, temporal smoothing is used in this approach to optimize the estimations based on the colour appearance model. Although this is a reliable method to automatically initialize a colour (specific) appearance model for an unknown person, the specific appearance model cannot be learned when the lateral walking pose does not exist in the video sequence, and the reliance on using a single frame to learn the colour appearance means the specific appearance model may not be robust with regard to lighting variations throughout the video sequence.

Ferrari *et al.* (2009) also seek to obtain an appearance model of the tracked person but do not use a stylized pose detector. The core of their estimation is an extension of the ‘image parsing’ approach of Ramanan (2006), which uses a generic parts detector based on edges and a pictorial structure to obtain an optimal pose estimation at each frame. The top few of these per-frame optimal estimations are then used to construct a specific (colour-based) appearance model as well as a colour model of the background and a second pose detection phase is then executed based on applying the pictorial structure to the average between the probability maps of the generic and specific detection models. In addition to this core approach, Ferrari *et al.* (2009) utilises a complex mix of supplementary methods such as upper body detection via histograms of oriented gradients, foreground highlighting and loopy temporal smoothing to improve the final detections. The advantages of this method are that it does not require a stylized pose to appear. Rather it utilizes evidence from both generic and specific detectors, and it samples colours from multiple frames and hence includes lighting variations into the specific appearance model. However, the effectiveness of the specific appearance model depends on the assumption that the generic detection likelihood is representative of accuracy *between* frames and so that the top few optimal estimations will be the most accurate. This assumption was not verified by Ferrari *et al.* (2009) – the issue here is that it is comparing pose estimations that have been extracted from different data (different frames), hence it is not certain that detection accuracy and detection likelihood will correlate *between* frames. Moreover, their choice to merge the generic and specific appearance maps via averaging does not ensure that both models must agree on the final estimation. For example, a very high likelihood in colour and low likelihood in generic can still produce a reasonable estimation in their method.

2.5.2 Specific Appearance Model vs Generic Appearance Model

As mentioned above, most pictorial structure based approaches are typically designed for estimating pose for various persons with various appearances. A generic appearance model

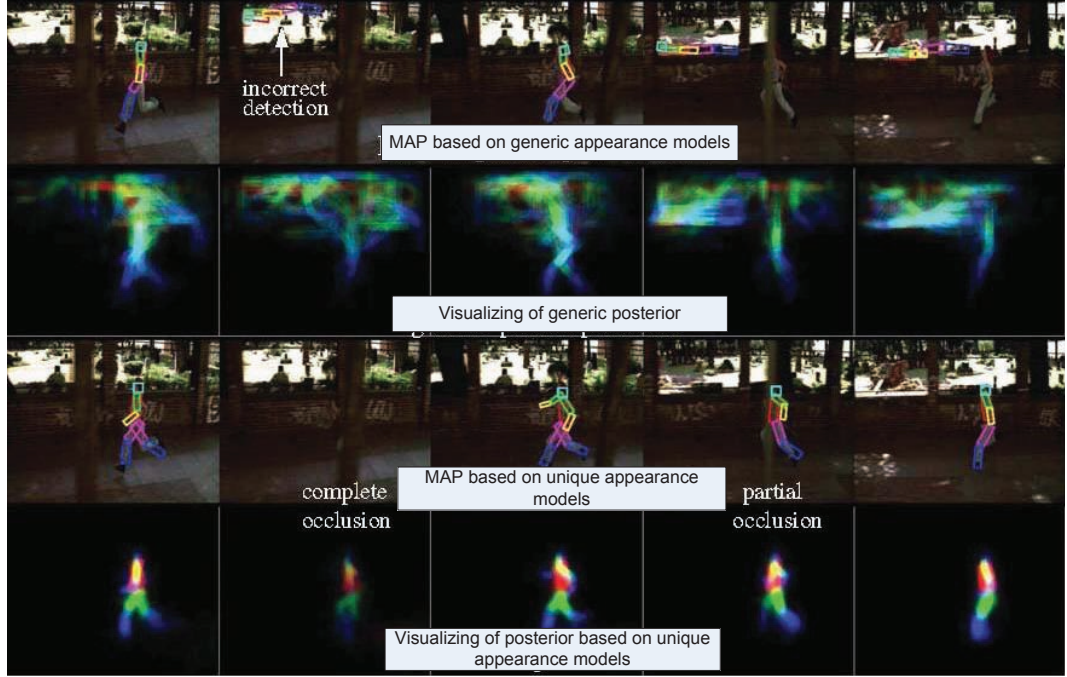


Figure 2.7: An example from [Ramanan *et al.* \(2007\)](#) is used to demonstrate that in comparison with a generic appearance model a specific appearance model facilitates the removal of interference from background clutter in estimating human pose.

is thus needed to describe the common appearance features of a human figure, which can generalize across all possible people wearing all possible clothes. The accuracy of such approaches in estimating human pose is generally not high due to the fact that a generic appearance model is usually built upon generic features (such as shape) so that it is easily confused by background clutter, causing spurious modes in its posterior. When the specific features (such as different colours of clothes in different body parts) of the tracked person are available, a more discriminative appearance model can be built based on these specific features, which is more effective in eliminating background clutters and distinguishing the human body part from the background than a generic appearance model. Thus, a higher accuracy in estimating human pose can be achieved. Here an example from [Ramanan *et al.* \(2007\)](#) is used to demonstrate the above discussion. In Figure 2.7, human pose estimations by applying a generic appearance model in pictorial structure are shown in the top two rows, which include both the MAP estimations (first row) and the visualizing of the entire posterior estimations (second row). In the second-row images, we can clearly see that many background clutters are identified by the generic appearance model as candidates thus resulting in incorrect pose estimations shown in the first-row images. In contrast, human pose estimations by applying a specific appearance model in pictorial structure are shown in the bottom two rows. The fourth-row images show that the specific appearance

model effectively removes the interference from the background clutters, thereby obtaining more accurate pose estimations as shown in the third row.

2.5.3 Weakness of the Approaches based on Specific Appearance Model

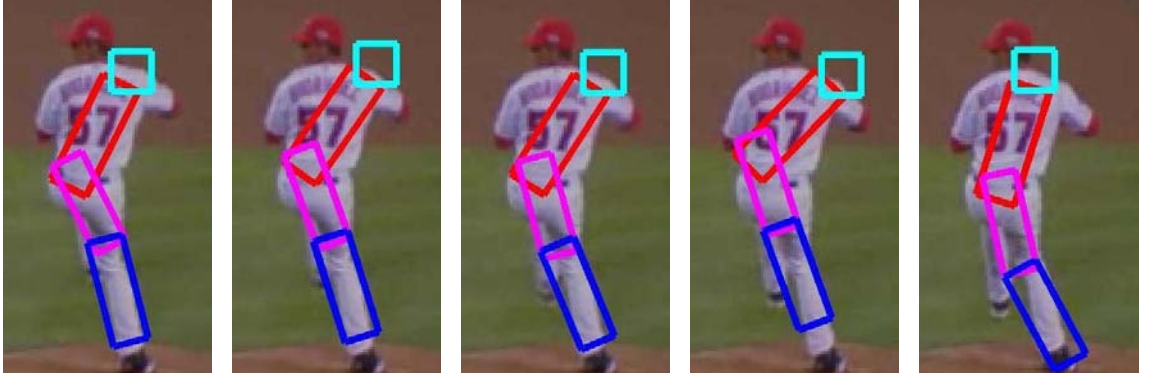


Figure 2.8: To describe the drawbacks of approaches only based on colour appearance model, some parts of Ramanan’s results are presented. Since the upper arms has similar colour with the torso and the lower arms are hidden by the torso, the system is unable to detect where the upper arms are.

As established in the previous section, a generic appearance model can be easily confused by background clutters due to the weak discriminative power of low-level features. However a specific appearance model may also be confused by some part of the foreground due to the choice of appearance features. For example, in [Ramanan *et al.* \(2007\)](#), the specific appearance models are built based on colour information. In many videos where the human wears a uniformly-coloured short-sleeved shirt, the upper arms appear to have the same colour as the torso, while the lower arms appear in a different colour. For such video sequences the appearance model for the upper arms is easily confused by the colour appearance of the torso. As a result, the estimated positions of the lower arms must be used to determine the configurations of the upper arms, not the other way round. As the lower arms are often occluded by the torso or other body parts, the configurations of the upper arms will fail to be estimated in such circumstances regardless of whether the upper arms are actually visible. Figure 2.8 shows such an example from [Ramanan *et al.* \(2007\)](#). It shows that when the lower arms are occluded by the torso and the upper arms have the same colour as the torso, the configurations of the upper arms are unable to be estimated in the system that relies only on a specific appearance model based on colour.

In addition, a specific appearance model often leads to ambiguities in estimating the orientation of human body parts, especially for torso. Several inaccurate estimations

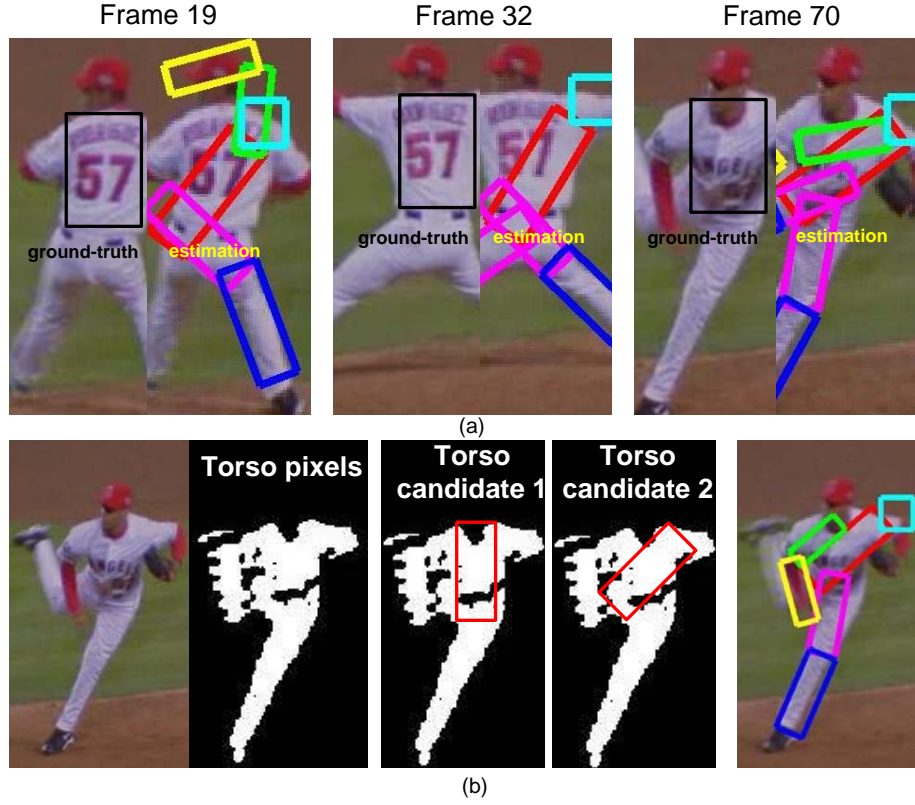


Figure 2.9: Figure (a) provides several inaccurate estimations for torso orientation from [Ramanan *et al.* \(2007\)](#). Figure (b) provides a demonstration on why a specific appearance model confuses the orientation estimation.

for torso orientation from [Ramanan *et al.* \(2007\)](#) are shown in Figure 2.9. In Figure 2.9 (a), the consistency in the torso orientations is poor between the estimates and the corresponding ground-truth. Figure 2.9 (b) provides some analysis on why a specific appearance model confuses the orientation estimation. Given that the colour model of the torso is known for the baseball player, the binary image for the torso pixels can be derived using the torso colour model where the white and black pixels represent torso and background pixels respectively. Many candidates exist for the torso estimation. Only two of them are drawn here in the binary image for our discussion. Each of the two candidates covers a local image region represented by a rectangle. The appearance likelihood of each candidate is evaluated by summing up the number of non-torso pixels inside its corresponding rectangle. The more non-torso pixels a candidate consists of, the lower value the appearance likelihood the candidate achieves. The correctness of the torso orientation is not taken into consideration here since it is not represented in the specific appearance model. Comparing the two candidates shown here, it is evident that the candidate with better appearance likelihood value actually corresponds to an incorrect pose, but it will be chosen since it has better appearance likelihood. An inaccurate torso estimation is hence

produced. Beside the ambiguities in estimating the orientation of human body parts, a specific appearance model could also possibly result in wrong estimations when two limbs are not well separated, as shown in Figure 2.10 . The incorrect estimate is caused by the unsatisfactory background subtractions and the consequent good appearance likelihood obtained for pixels located between the two lower legs.

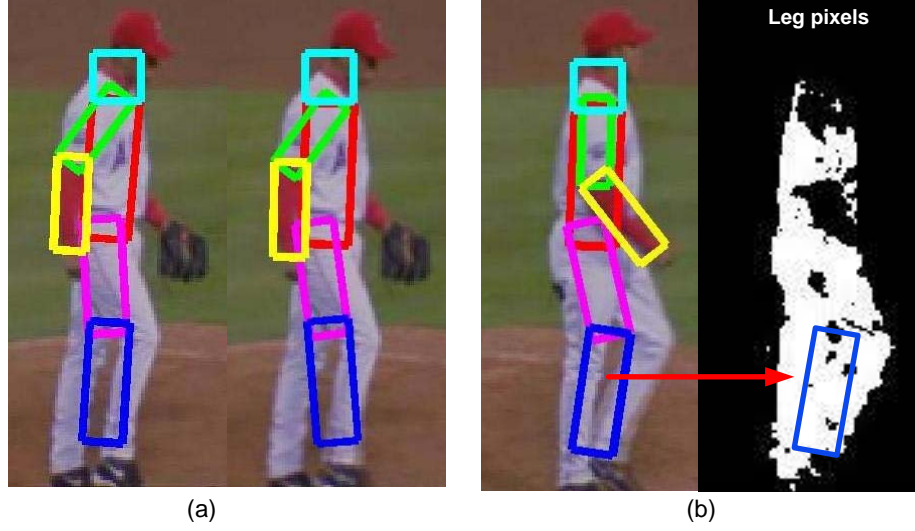


Figure 2.10: Some examples are provided to show that a specific appearance model can result in wrong estimations when two limbs are not well separated.

2.5.4 Section Summary

In this section, two state-of-art approaches ([Ramanan *et al.*, 2007](#); [Ferrari *et al.*, 2009](#)) that are most relevant to this thesis are reviewed and their weaknesses are analyzed. We also analyze the weaknesses of the approaches based on a generic appearance model in comparison with the approaches based on a specific appearance model. Finally, the weaknesses of the approaches based on a specific appearance model are analyzed.

2.6 Chapter Summary

The chapter has presented a review of existing work that is relevant to this thesis. It begins with a review of current methods for human finding and tracking. These methods are usually used as a pre-processing step in the systems for human pose estimation. Next model-free approaches for human pose estimation are presented, with their advantages

and disadvantages analyzed to justify why this type of approaches are not suitable for the objectives of accurately estimating human pose. Model-based approaches, which have been determined to be more suitable for the purpose of this thesis, are then described. The approaches are categorized as top-down and bottom-up approaches. A brief overview of top-down approaches shows that to accurately estimate human pose using the top-down approaches, at least one of two conditions (multi-views and motion-specified) needs to be satisfied, which is inconsistent to our objective (monocular view and arbitrary motion) in this thesis. Bottom-up approaches for human pose estimation are then reviewed with particular focus on the approaches under the framework of pictorial structure. It has been established that the bottom-up approaches with the framework of pictorial structure are suitable for our objectives. Two state-of-art approaches that are most relevant to this thesis are then reviewed, the work by [Ramanan *et al.* \(2007\)](#) and [Ferrari *et al.* \(2009\)](#). We aim to overcome the weaknesses of the existing works to build a system for to better estimate human with arbitrary motions in monocular images/videos.

The next chapter will provide a detailed description of the framework of pictorial structures and explore how the framework can be employed and enhanced to form the basis of a system for human pose estimation.

Chapter 3

Pictorial Structure Framework

As discussed in Chapter 2, the existing approaches for human body tracking have been briefly reviewed and analyzed. The work that are most closely related to the aims of this thesis are two state-of-art approaches (Ramanan *et al.*, 2007; Ferrari *et al.*, 2009) that have the same objective of estimating human pose over time by applying a specific (colour) appearance model of a tracked person via the pictorial structure. In order to automatically learn a specific appearance model for a tracked person, they both attempt to select the accurate training samples in the generic detections from a generic human pose detector, which is built based on the framework of pictorial structure.

Different methods are used in Ramanan *et al.* (2007) and Ferrari *et al.* (2009) to obtain the accurate training samples from the generic detections. In Ramanan *et al.* (2007), a stylized pose detector that is specific for estimating lateral walking pose is designed based on the framework of a pictorial structure. Even though this method can accurately estimate the lateral walking pose and thus obtain accurate training samples, this method has the obvious limitation that it can not be applied in sequences in which the specified human pose does not exist. In contrast to Ramanan *et al.* (2007), Ferrari *et al.* (2009) attempts to select the accurate training samples based on a generic human pose detector that can estimate arbitrary human pose. Based on the assumption that the accuracy of the optimal estimations between frames can be represented by their likelihood, the top few optimal estimations are selected as the training samples by comparing the likelihood of the optimal estimations between different frames. However, although the accuracy of generic detections can be reasonably represented by their likelihood within a frame. It is less clear whether the likelihood of generic detections from *different data* (from different frames) are comparable. Ferrari *et al.* (2009) relies on assuming that this is the case, but does not provide any experimental verification to support it.

In this chapter, we build a generic human pose detector based on the framework of pictorial structure and then utilize it to test the assumption of Ferrari *et al.* (2009). Then we further investigate the characteristics of the generic human pose detector by analyzing the optimal estimations and the sub-optimal estimations. However, in order to provide sufficient context to assist the reader in following our analysis and subsequent conclusions,

we first review in-depth the pictorial structure and its use in human pose estimation. Specifically, we focus on the statistical perspective of the pictorial structure and re-derive the algorithms for finding the optimal match and for sampling from the pose estimate posterior.

This chapter is organized as follows. Section 3.1 reviews the framework of the pictorial structure and some important algorithms under this framework. Section 3.2 describes how to build a generic human pose detector based on the framework of pictorial structure and the outputs of the generic human pose detector. Section 3.3 verifies the assumption of Ferrari *et al.* (2009) based on the results of generic human pose detector. This is followed by an investigation in Section 3.4 on the characteristics of the generic human pose detector, including analyzing the accuracy of the optimal estimations and sub-optimal estimations. Finally, a summary of the chapter is presented in Section 3.5.

3.1 Pictorial Structure Framework

A pictorial structure model for object representation, which was first introduced by Felzenszwalb and Huttenlocher (2005), is composed of a set of parts and connections between certain pairs of parts. It is a general framework because it does not rely on the specific fashion of modelling the appearance of each part nor the type of connections between parts. An undirected graph $G = (V, E)$ is usually used to describe such a model. The vertices $V = \{v_1, \dots, v_n\}$ correspond to n parts. The connection between parts v_i and v_j is represented by an edge $(v_i, v_j) \in E$. An instance of an object is specified by a configuration $L = (l_1, \dots, l_n)$ where each l_i specifies the location of part v_i .

The problem of matching a pictorial structure to an image is defined as minimizing an energy function, as can be achieved via the algorithm of Fischler and Elschlager (1973). The energy of a specific configuration is determined both by the degree of matching a body part to the image data at its position and the degree of agreement between the relative locations of the parts. Given an image, a function $m_i(l_i)$ is defined to measure the degree of mismatch between part v_i and the image data located at l_i in the image. For a given pair of connected parts v_i and v_j , a function $d_{ij}(l_i, l_j)$ is defined to measure the degree of deformation of the model when parts v_i and v_j are respectively located at l_i and l_j in the image. An optimal match of the model to the image is then naturally defined as

$$L^* = \arg \min_L \left(\sum_{i=1}^n m_i(l_i) + \sum_{(v_i, v_j) \in E} d_{ij}(l_i, l_j) \right), \quad (3.1)$$

which is a configuration minimizing the sum of the match costs m_i for each part v_i and

the deformation costs d_{ij} for connected pairs of parts. The model is invariant to certain global transformations, as the deformation cost is generally defined as a function of the relative locations between two connected parts. An important characteristic in the process of matching the pictorial structure model to an image is that the process of matching does not alone decide any individual part's position. Rather it makes an overall decision for all parts together based on both the part matching costs and the deformation costs.

In order to find a global minimum of the energy function by an efficient matching algorithm, such as the belief propagation algorithms ([Pearl \(1988\)](#)), two restrictions are required to be enforced on the pictorial structure model. First, the graph G is required to be acyclic which means that the graph is tree-structured. The connections between human body parts can be naturally represented by a tree structure in that human skeletal structure can be naturally transformed to a tree structure with torso as the root. Second, the deformation cost between a connected pair of parts is required to be represented in a particular form where $d_{ij}(l_i, l_j)$ is required to be a Mahalanobis distance between the transformed locations,

$$d_{ij}(l_i, l_j) = \left(T_{ij}(l_i) - T_{ji}(l_j) \right)^T M_{ij}^{-1} \left(T_{ij}(l_i) - T_{ji}(l_j) \right), \quad (3.2)$$

where the matrix M_{ij} is required to be diagonal and functions T_{ij} and T_{ji} are one-to-one functions. This restriction imposed on the form of connections between parts makes the running time of matching algorithm become linear with respect to the number of possible positions for each part ([Felzenszwalb and Huttenlocher, 2005](#)).

To apply a pictorial structure model for finding a particular object in an image, a corresponding model has to be constructed. The model is required to define the appearance parameters for each part, a set of edges connecting pairs of parts and the characteristics of the connections. The problem to specify a good model for a particular object is crucial for accurately finding the configuration of the particular object. If these parameters in the model can be learned automatically from training data, the model will be more practical. As discussed above, the problem of matching a pictorial structure model to an image is characterized as the minimization of an energy function, which only allows for finding the best match, whereas it is often desirable to find multiple good potential matches. Fortunately, the energy minimization problem can be transformed to a problem of statistic estimation under a statistical formulation ([Felzenszwalb and Huttenlocher, 2005](#)), which makes it possible for the parameters of pictorial structure model to be learned from training data and enable finding multiple good matches of a structure model in an image.

3.1.1 Statistical Framework of Pictorial Structure Model

The statistical framework of pictorial structure is helpful for addressing two problems, learning the parameters of a model automatically from training data and finding multiple good potential matches of a model to an image.

The problem of matching a pictorial structure model to an image can be described in a statistical formulation as follows. Let θ denote a set of parameters which defines the pictorial structure model. Let I represent an image and as before let L denote the configuration of an object that consists of the locations of all the parts. The distribution $P(I|L, \theta)$ characterizes the imaging process and measures the likelihood of seeing a particular image given that an object is at a particular configuration. The distribution $P(L|\theta)$ measures the prior probability given that the object is in a particular configuration. Finally, the posterior distribution $P(L|I, \theta)$ captures the probability that the object is at a particular configuration L given the image I and the model parameters θ . Based on Bayes' rule the posterior distribution can be expressed as

$$P(L|I, \theta) \propto P(I|L, \theta)P(L|\theta). \quad (3.3)$$

Three important problems can be described in terms of this statistical formulation of the pictorial structure model,

1. The optimal match of a pictorial structure model to an image is defined as finding a configuration L to get the maximum a posterior (MAP) probability in the posterior distribution $P(L|I, \theta)$, which essentially uses a MAP estimation so as to get the best guess for the configuration of the object. In the pictorial structure framework, the MAP estimation will be equivalent to the estimation of the energy minimization which is defined in Equation (3.1).
2. The problem of finding multiple good potential matches of a model to an image can be transformed to the problem of sampling from the posterior distribution. The sampling provides a natural way to select multiple good potential matches of a model to an image. This characteristic inspires us to propose an approach where the generic appearance model of the human and the specific appearance model of the tracked person are naturally combined together for the estimation of a human pose, which will be discussed in Chapter 4.
3. In this statistical formulation, the parameters for the pictorial structure model can be learned through the use of maximum likelihood estimation.

In this thesis, the parameterization of [Felzenszwalb and Huttenlocher \(2005\)](#) is used, specifically $\theta = (E, c, u)$, where E denotes a set of edges that specify which pairs of parts are connected, $c = \{c_{ij} | (v_i, v_j) \in E\}$ denote the connection parameters and $u = \{u_1, \dots, u_n\}$ denote the appearance parameters.

3.1.1.1 Approximated Appearance Likelihood Model

In practice, the likelihood distribution $p(I|L, \theta)$ from Equation (3.3) is approximated by the product of all individual likelihoods,

$$p(I|L, \theta) = P(L|I, u) \propto \prod_{i=1}^n p(I|l_i, u_i), \quad (3.4)$$

where the individual likelihood distribution $p(I|l_i, u_i)$ measures the likelihood of seeing an image I given that the configuration of a part is l_i while the appearance of the part is u_i . This is a good approximation if there are no overlapping parts appearing in the configuration of the object. However, for articulated objects such as human where less constraints are inherent in the locations of parts, parts can easily overlap. In this case, a configuration estimation based on this likelihood may become poor. This problem is possibly handled by generating multiple samples from the posterior distribution and choosing a good estimate from them based on an independent method.

3.1.1.2 Tree-structured Model

The distribution $P(L|\theta)$ from Equation (3.3) is captured by a tree-structured Markov random field with edge set E . The distribution for a tree-structured prior can be expressed as ([Felzenszwalb and Huttenlocher, 2005](#)):

$$p(L|\theta) = p(L|E, c) \propto \prod_{(v_i, v_j) \in E} p(l_i, l_j | c_{ij}). \quad (3.5)$$

where v_i and v_j are vertices in the graph defined by E .

3.1.1.3 Approximated Posterior Distribution

In Equation (3.4), the distribution $P(I|L, \theta)$ measures the likelihood of seeing an image given the particular configuration of an object and its appearance parameters. This is

approximated by the product of individual likelihoods. Similarly, in Equation (3.5), the tree-structured prior distribution $p(L|\theta)$ is decomposed into the product of individual prior distributions between the connected parts. They both can be substituted back into Equation (3.3) yielding,

$$p(L|I, \theta) \propto \left(\prod_{i=1}^n p(I|l_i, u_i) \prod_{(v_i, v_j) \in E} p(l_i, l_j | c_{ij}) \right). \quad (3.6)$$

This can be placed in the same form as the energy function that is being minimized in Equation (3.1) by taking the negative logarithm,

$$-\log p(L|I, \theta) \propto \left(\sum_{i=1}^n -\log p(I|l_i, u_i) + \sum_{(v_i, v_j) \in E} -\log p(l_i, l_j | c_{ij}) \right). \quad (3.7)$$

Comparing Equation (3.7) with the energy function that is minimized in Equation (3.1), it is shown that the negative logarithm of the posterior is the energy function where $m_i(l_i) = -\log p(I|l_i, u_i)$ that measures the degree of mismatch between part v_i and the image data located at l_i and $d_{ij} = -\log p(l_i, l_j | c_{ij})$ that measures the degree of deformation of the model when parts v_i and v_j are respectively located at l_i and l_j . Thus it is easy to understand that the problem of MAP estimation in the statistical model is identical to the problem of energy minimization for a pictorial structure.

As discussed in the beginning of Section 3.1, in order to find a global minimum of the energy function by an efficient matching algorithm, the deformation costs d_{ij} are required to be expressed in a particular form as shown in Equation (3.2). This requirement has a corresponding expression in terms of the statistical models. Since $d_{ij} = -\log p(l_i, l_j | c_{ij})$, it is equivalent to assume that the prior distribution $p(l_i, l_j | c_{ij})$ is given by a Gaussian over the displacement between transformed locations,

$$p(l_i, l_j | c_{ij}) \propto \mathcal{N}(T_{ij}(l_i) - T_{ij}(l_j), 0, D_{ij}), \quad (3.8)$$

where T_{ij} , T_{ji} and D_{ij} are the connection parameters represented by c_{ij} . These parameters correspond to the parameters in Equation (3.2) where $D_{ij} = M_{ij}/2$ is a diagonal covariance matrix.

3.1.2 Algorithm for Finding the Optimal Match

The problem of finding the best match of a pictorial structure model to an image can be characterized as Equation (3.1) restated here for the readers' convenience.

$$L^* = \arg \min_L \left(\sum_{i=1}^n m_i(l_i) + \sum_{(v_i, v_j) \in E} d_{ij}(l_i, l_j) \right).$$

Felzenszwalb and Huttenlocher (2005) restrict the graph $G = (V, E)$ to a tree form, making it feasible to apply a polynomial-time dynamic programming algorithm (described in Section 3.1.2.1). From this algorithm, a more efficient minimization approach can be applied using the generalized distance transform as developed by Felzenszwalb and Huttenlocher (2005) and briefly described in Section 3.1.2.2.

3.1.2.1 Dynamic programming for Efficient Minimization

In this section, an algorithm developed by Felzenszwalb and Huttenlocher (2005) is described for finding a configuration $L^* = \{l_1^*, \dots, l_n^*\}$ which minimizes Equation (3.1) when the graph is a tree. Given $G = (V, E)$, let $v_r \in V$ denote an arbitrarily selected root vertex. The results will not be affected by this selection. Let d_i be the depth of the vertex v_i , which is the number of edges between v_i and v_r (the depth of v_r is 0). The children C_i of v_i is defined as the vertices of depth $(d_i + 1)$ which are connected to v_i , if any. Every vertex other than the root vertex has a unique parent, which is the neighbouring vertex with depth $(d_i - 1)$.

For any vertex v_j without children, the best location l_j^* can be represented by a function of its unique parent v_i . The only edge incident on v_j is (v_i, v_j) , so that the only contribution of v_j to the energy function described in Equation (3.1) is $m_j(l_j) + d_{ij}(l_i, l_j)$. The best contribution of v_j to the energy function given location l_i for v_i , $B_j(l_i)$, is expressed as

$$B_j(l_i) = \min_{l_j} (m_j(l_j) + d_{ij}(l_i, l_j)), \quad (3.9)$$

and the best location for v_j as a function of v_i is expressed as

$$B_j^*(l_i) = \arg \min_{l_j} (m_j(l_j) + d_{ij}(l_i, l_j)). \quad (3.10)$$

For any vertex v_j except for the root vertex, assume that $B_c(l_j)$ is known for each child vertex $v_c \in C_j$, which means that the best contribution of each child vertex v_c to the energy function given the location l_j for v_j is known. The best contribution of v_j to the energy function as a function of his parent v_i is expressed as

$$B_j(l_i) = \min_{l_j} \left(m_j(l_j) + d_{ij}(l_i, l_j) + \sum_{v_c \in C_j} B_c(l_j) \right), \quad (3.11)$$

and the best location of v_j with respect to the location l_i for v_i can be obtained from

$$B_j^*(l_i) = \arg \min_{l_j} \left(m_j(l_j) + d_{ij}(l_i, l_j) + \sum_{v_c \in C_j} B_c(l_j) \right). \quad (3.12)$$

Finally, for the root v_r , assume that $B_c(l_r)$ for each child vertex $v_c \in C_r$ then the best location for the root vertex is

$$l_r^* = \arg \min_{l_r} \left(m_r(l_r) + \sum_{v_c \in C_r} B_c(l_r) \right). \quad (3.13)$$

The minimization in Equation (3.1) can then be transformed to a recursive form in terms of $(n - 1)$ functions $B_j(l_i)$ for each vertex $v_j \in V$ (except for the root vertex). An efficient minimization algorithm is suggested by these recursive equations. Assume that the maximum depth is d in the tree, the algorithm composes of three steps,

1. For each vertex v_j with $depth = d$, $B_j(l_i)$ is computed according to Equation (3.9) and $B_j^*(l_i)$ is computed as in Equation (3.10) where v_i is the parent of v_j .
2. For each vertex v_j with $depth = (depth - 1)$, each $B_c(l_j)$ where $v_c \in C_j$ has been computed in previous step, thus $B_j(l_i)$ is computed in terms of Equation (3.11) and $B_j^*(l_i)$ is computed as in Equation (3.12).
3. Repeat the second step until $depth = 0$.

After each $B_j(l_i)$ for each vertex $v_j \in V$ (except for the root vertex) has been computed, the optimal location l_r^* for the root can be computed using Equation (3.13). The optimal location L^* for all the parts can be obtained by tracking from the root to each leaf. The optimal location for each vertex can be computed given the optimal location of its parent vertex. Starting from the optimal location of the root, the optimal location of each parent vertex can be obtained.

3.1.2.2 Generalized Distance Transforms

When d_{ij} is restricted in the form of Equation (3.2), in order to compute each $B_j(l_i)$ more efficiently, Felzenszwalb and Huttenlocher (2005) introduced the concept of the generalized distance transform. It differs from a traditional distance transform in that it measures distances between sets of arbitrarily-located points rather than for sets of points arranged on a grid. First the traditional distance transform will be reviewed before introducing the generalization.

Given a grid \mathcal{G} , $\rho(x, y)$ is a certain measure of distance between points on the grid. Given a point set $B \subseteq \mathcal{G}$, the distance transform based on the set B specifies a value for each

point in the grid \mathcal{G} . For any x in the \mathcal{G} , the specified value $\mathcal{D}_B(x)$ is defined as the distance to the closest point in the set B , which is expressed as,

$$\mathcal{D}_B(x) = \min_{y \in B} (\rho(x, y)).$$

For efficient computation of the distance transform, the distance transform is commonly expressed in another form,

$$\mathcal{D}_B(x) = \min_{y \in \mathcal{G}} (\rho(x, y) + 1_B(y)),$$

where

$$1_B(y) = \begin{cases} 0 & \text{if } y \in B \\ \infty & \text{otherwise.} \end{cases}$$

This suggests a generalization of the distance transforms where the indicator function $1_B(y)$ is replaced with some arbitrary function $f(y)$ over the grid \mathcal{G} ,

$$\mathcal{D}_f(x) = \min_{y \in \mathcal{G}} (\rho(x, y) + f(y)).$$

With the restricted form of d_{ij} in Equation (3.2), the functions $B_j(l_i)$ defined in Equation (3.11) can be rewritten as generalized distance transforms,

$$B_j(l_i) = \mathcal{D}_f(T_{ij}(l_i)),$$

where

$$f(y) = \begin{cases} m_j(T_{ij}^{-1}(y)) + \sum_{v_c \in C_j} B_c(T_{ji}^{-1}(y)) & \text{if } y \in \text{range}(T_{ji}) \\ \infty & \text{otherwise.} \end{cases}$$

and the distance in the grid, $\rho(x, y)$, is given by the Mahalanobis distance defined by M_{ij} ,

$$\rho(T_{ij}(l_i), T_{ji}(l_j)) = \left(T_{ij}(l_i) - T_{ji}(l_j) \right)^T M_{ij}^{-1} \left(T_{ij}(l_i) - T_{ji}(l_j) \right).$$

The grid \mathcal{G} denotes a discrete set specifying all possible values for $T_{ji}(l_j)$ that are considered during the minimization. However, the discrete set for $T_{ji}(l_j)$ and the discrete set for $T_{ij}(l_i)$ normally cannot coincide. Thus, an approximation is made by defining the value of the distance transform at a non-grid position to be the value at the closest grid point.

3.1.3 Algorithm for Sampling from the Posterior

After discussing the problem of finding the best match of a pictorial structure to an image, the algorithm for sampling from the posterior is discussed in this section. This algorithm

is used for finding multiple good potential matches of a pictorial structure model to an image, which is crucial for our proposed approaches.

As described in Section 3.1.1.3 the approximated posterior distribution for tree-structure pictorial structure models is given by

$$p(L|I, \theta) \propto \left(\prod_{i=1}^n p(I|l_i, u_i) \prod_{(v_i, v_j) \in E} p(l_i, l_j | c_{ij}) \right).$$

As described before, $v_r \in V$ denotes the root vertex (in the case of this thesis, this will be the torso) and C_i denotes the children of v_i . A sampling algorithm works by first computing $p(l_r, |I, \theta)$ which is the posterior distribution for the root vertex and then sampling a location for the root. Given this sampled location l_r , a location for each child, v_c , of the root is sampled from $p(l_c | l_r, I, \theta)$. This procedure is repeated until locations for all parts have been sampled. The marginal distribution for the root location is,

$$p(l_r | I, \theta) \propto \sum_{l_1} \dots \sum_{l_{r-1}} \sum_{l_{r+1}} \dots \sum_{l_n} \left(\prod_{i=1}^n p(I|l_i, u_i) \prod_{(v_i, v_j) \in E} p(l_i, l_j | c_{ij}) \right).$$

Using the independence assumptions between parts implied by the tree structure, this distribution can be rewritten as

$$p(l_r | I, \theta) \propto p(I | l_r, u_r) \prod_{v_c \in C_r} S_c(l_r). \quad (3.14)$$

The functions $S_j(l_i)$ are analogous to the $B_j(l_i)$ used for the energy minimization algorithm,

$$S_j(l_i) \propto \sum_{l_j} \left(p(I | l_j, u_j) p(l_i, l_j | c_{ij}) \prod_{v_c \in C_j} S_c(l_j) \right). \quad (3.15)$$

A polynomial algorithm to compute $p(l_r | I, \theta)$ can be obtained from the above recursive functions. A more efficient algorithm for computing each $S_j(l_i)$ in the case where $p(l_i, l_j | c_{ij})$ is in the special form given by Equation (3.8) will be described in Section 3.1.3.2. As with the computation process in the energy minimization algorithms, the S functions can be computed by starting from the leaf vertices (Felzenszwalb and Huttenlocher, 2005).

3.1.3.1 Sampling from the posterior

After the S functions for every part except for the root are obtained, the marginal distribution $p(l_r|I, \theta)$ will be computed by Equation (3.14). A location for the root can then be sampled from $p(l_r|I, \theta)$. Next we need to consider how to sample the locations for other parts. Given a sampled location l_i for v_i , the posterior for the child v_j of v_i can be expressed as

$$p(l_j|l_i, I, \theta) \propto p(I|l_j, u_j)p(l_i, l_j|c_{ij}) \prod_{v_c \in C_j} S_c(l_j). \quad (3.16)$$

Since all the S functions are available now, the posterior distribution $p(l_j|l_i, I, \theta)$ can be computed immediately. Thus once a location for the root has been sampled, a location can be sampled for each of its children. The sampling process can be continued in this manner until a location is sampled for every part. It is important to note that multiple sampling will not take much more time than single sampling in that the program only needs to compute the S function once. This will be utilised in Section 3.4.2 to sample multiple alternative estimations.

3.1.3.2 An efficient algorithm for computing the S functions

To efficiently compute the functions in Equation (3.15), they are written as a Gaussian convolution in the transformed space of locations given by T_{ij} and T_{ji} . Using the special form of $p(l_i, l_j|c_{ij})$, Equation (3.15) can be written as

$$S_j(l_i) \propto \sum_{l_j} \left(\mathcal{N}(T_{ij}(l_i) - T_{ji}(l_j), 0, D_{ij}) p(I|l_j, u_j) \prod_{v_c \in C_j} S_c(l_j) \right).$$

Felzenszwalb and Huttenlocher (2005) shows that this can be interpreted as the Gaussian convolution in the transformed space:

$$S_j(l_i) \propto (F \otimes f)(T_{ij}(l_i)),$$

where F is a Gaussian filter with covariance D_{ij} , \otimes is the convolution operator, and

$$f(y) = \begin{cases} p(I|T_{ji}^{-1}(y), u_j) + \prod_{v_c \in C_j} S_c(T_{ji}^{-1}(y)) & \text{if } y \in \text{range}(T_{ji}) \\ 0 & \text{otherwise.} \end{cases}$$

This convolution is done over a discrete grid which specifies possible values for $T_{ji}(l_j)$. The Gaussian filter F is separable since the covariance matrix is diagonal. A good approximation for the convolution can be computed in time linear in h' using the techniques from Wells (1986). This gives an overall $O(h'n)$ time algorithm for sampling a configuration from the posterior distribution. The sampling algorithm for posterior used in this thesis follows the above-mentioned algorithm.

3.2 Generic Human Pose Detector

In previous sections, we introduce the original pictorial structure framework and the related algorithms. Such a general framework can be used to build a pose detector to estimate the configuration of various joint-connected objects such as animals and human. We only focus on estimating human configuration in this thesis. A variety of human pose detectors has been proposed under this framework and they are different from each other in the appearance likelihood and configuration prior models. In this section, a generic human pose detector (generic detector) used in this thesis is defined based on the pictorial structure framework. Specifically, the configuration prior and appearance likelihood model will be defined and we will discuss what outputs can be generated by this generic detector.

In our generic pose detector, the human configuration is denoted as a set $L = \{l_n\}$, $n \in \{\text{head, torso, left-upper-arm, right-upper-arm, left-lower-arm, right-lower-arm, left-upper-leg, right-upper-leg, left-lower-leg, right-lower-leg}\}$, where $l_n = (x_n, y_n, \theta_n)$ represents the position (x_n, y_n) and orientation θ_n of part n in an image. The appearance model is denoted as a set $C = \{C_n\}$, where C_n is the appearance model for part n . Denoting the image data as I , the posterior of the human body configuration L can be written as:

$$P(L|I, C) \propto P(L)P(I|L, C), \quad (3.17)$$

where $P(I|L, C)$ is the appearance likelihood for the given configuration L and the given appearance model C , i.e., $P(I|L, C)$ is the possibility of the image representing the human with configuration L and appearance model C , and $P(L)$ represents the geometric relations between the connected parts, and thus also called the configuration prior.

3.2.1 The Configuration Prior

The first component in the pictorial structure is the configuration prior $P(L)$, which represents the constraints between the connected parts in the probabilistic form. We use the kinematical dependencies between parts as the constraints. The kinematical dependency relations in the human body can be represented as a directed acyclic graph as shown in Figure 3.1.

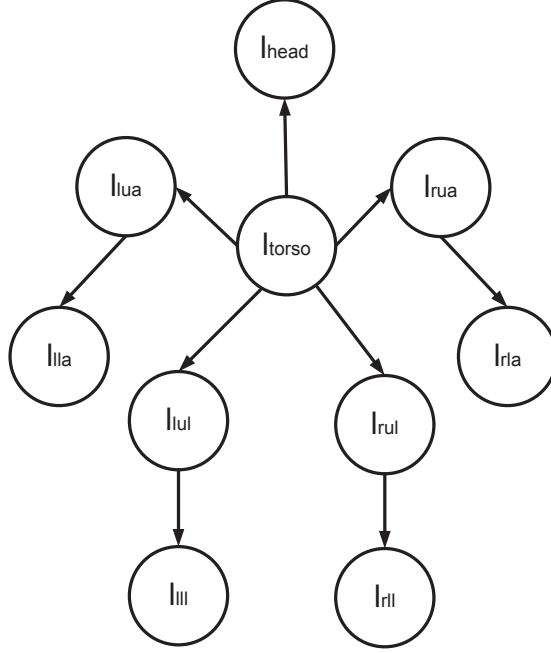


Figure 3.1: Kinematic tree model. Each node represents a body part. The root node is denoted by l_{torso} . The subscripts denote the body parts. In the abbreviation of subscripts, the first character l or r denotes right or left. The second character u or l denotes upper or lower. The last character a or l denotes arm or leg. The dependency relations are denoted by directed edges.

The configuration prior is organized in a kinematical tree, so that the distribution over configuration can be factorized as

$$P(L) = P(l_0) \prod_{(m,n) \in E} P(l_m | l_n). \quad (3.18)$$

In Equation (3.18), E is the set of body part connections, including the relations between head, torso, left/right upper/lower legs/arms in the kinematical tree, l_0 denotes the root node in the tree, which is the torso in the pictorial structure for a human body, each l_m is the body part and is only dependent on its immediate parent l_n . The prior for the root

node configuration $P(l_0)$ is assumed to be uniform, which avoids placing limits on the root configuration and hence allows for any body posture.

For efficient inference, the relation between parts $P(l_m|l_n)$ can be modelled as a Gaussian distribution. This may seem to place a significant limitation on representing the relations between parts, such as the relative orientation of lower arm given with respect to the upper arm is intuitively a semi-circular rather than Gaussian shape. However, it is possible to transform the configuration coordinates to a coordinate system of the body joint, to ensure that the spatial relations between connected parts are well captured by the Gaussian distribution. In particular, to model $P(l_m|l_n)$, the part configuration $l_m = (x_m, y_m, \theta_m)$ is transformed to the coordinate system of the joint between part m and part n by the transformation:

$$T_{nm}(l_m) = \begin{bmatrix} x_m + d_x^{nm} \cos \theta_m - d_y^{nm} \sin \theta_m \\ y_m + d_x^{nm} \sin \theta_m + d_y^{nm} \cos \theta_m \\ \theta_m + \tilde{\theta}_{mn} \end{bmatrix}, \quad (3.19)$$

where $d^{nm} = (d_x^{nm}, d_y^{nm})^T$ is the mean relative position of the joint between part m and part n in the coordinate system of part m and $\tilde{\theta}_{mn}$ is the relative angle between the two parts. The $P(l_m|l_n)$ can then be modelled as a Gaussian in the transformed space:

$$P(l_m|l_n) = N(T_{nm}(l_m)|T_{mn}(l_n), \sigma^{nm}) \quad (3.20)$$

where T_{mn} is the transformation that maps the position of part l_n to the position of the joint between part m and part n , and σ^{nm} is the covariance between the parts.

The spatial relation $P(l_m|l_n)$ can be automatically generated by learning d^{nm} and σ^{nm} using the maximum likelihood estimation. In this thesis the multi-view and multi-articulation people dataset provided by [Toyama and Blake \(2002\)](#) is used, where the human activities vary in a large range from a simple walking to performing acrobatic exercises.

3.2.2 The Appearance Likelihood Model

Another important component in the pictorial structure is $P(I|L, C)$, the appearance likelihood model for a particular body configuration. For simplicity, I_m , which is the image evidence of part m , is assumed conditionally independent given the body configuration L and appearance model C . Each I_m only depends on its own configuration l_m and the appearance model C_m . So the likelihood $P(I|L, C)$ can be simplified as

$$P(I|L, C) = \prod_{m=0}^N P(I_m|l_m, C_m). \quad (3.21)$$

Under our generic detector, C is required to be a generic appearance model that describes the common features of a human. To achieve accurate estimations, building a discriminative generic appearance evaluation model is important. In our detector, the appearance model introduced in [Andriluka *et al.* \(2009\)](#) is adopted. Our generic part detectors densely sample a variant of the shape context descriptor ([Mikolajczyk and Schmid, 2005](#); [Seemann *et al.*, 2005](#)). In this descriptor, the distribution of locally normalized gradient orientations is captured in a log-polar histogram. In our experiments we use 12 bins for the location and 8 bins for the gradient orientation, which results in a 96 dimensional descriptor. The sign of the gradient is ignored as it is found to improve generalization. The feature vector describing shape features of a particular body part is generated by concatenating image features whose centres fall inside the bounding box of a body part. During detection all possible positions, scales, and orientations are scanned in a sliding window fashion. An AdaBoost ([Freund and Schapire, 1995](#)) classifier is then trained in order to obtain a discriminative generic appearance evaluation model.

3.2.3 Outputs of the Generic Pose Detector

Substituting Equation (3.18) and Equation (3.21) into Equation (3.17), the posterior of the human body configuration L can be written as:

$$P(L|I, C) \propto P(l_0) \prod P(l_m|l_n) \prod P(I_m|l_m, C_m). \quad (3.22)$$

After the posterior of the human body configuration is generated, in most cases where human body parts is represented by a tree-structured graph and torso is chosen as the root node (such as the case in this section), the algorithm described in Section 3.1.2 is directly used to obtain the globally optimal estimation of human body configuration. A problem of this algorithm is that an incorrect optimal estimation of a parent node will unavoidably result in incorrect optimal estimations of the children nodes. This is because in this algorithm the optimal estimation of the children nodes are computed given the optimal estimation of its parent node. To avoid this problem, the exact marginal posterior for each body part is calculated in this thesis and then the estimation with maximum marginal posterior in each body part is chosen as the optimal estimation for that body part.

In our system, the generic detector will be applied to estimate the human configuration in each frame of a video sequence. In each frame, an optimal estimation for part m is first obtained, which is denoted by Q_m . The optimal estimations for part m in all frames

of the sequence are denoted as a set S_m , where $S_m = \{Q_m^1, Q_m^2, \dots, Q_m^t \dots\}$ and t is the index of image frames in the sequence.

We know that the imprecise models and background clutter may result in the optimal estimation not being the correct one. In such cases, it is important to explore whether the correct estimation is potentially hidden in these sub-optimal estimations. If there is evidence that the correct estimation is highly likely to be found in these sub-optimal estimations, an attempt can be made to obtain the correct estimation in these sub-optimal estimations. Therefore, besides the optimal estimation, the multiple good estimations (sub-optimal estimations) are obtained using the sampling algorithm described in Section 3.1.3, which is another output of the generic detector. Please note that the sampling algorithm will not take extra computational time. In fact, it is to choose a set of estimations in a set of good estimations found during the search for the best estimation. Formally, besides the optimal estimation for part m in a frame, a set of sub-optimal estimations for part m , denoted as U_m , is obtained by sampling the top N marginal posterior $P(l_m|I, C)$. These sub-optimal estimations are referred to as multiple probable alternative estimations.

3.3 Analysing the Assumption of Ferrari

In order to automatically learn a specific appearance model, [Ferrari *et al.* \(2009\)](#) proposed an approach to refine the optimal estimations in different frames based on their likelihood scores. More specifically, [Ferrari *et al.* \(2009\)](#) ranks the per-frame optimal estimations by their likelihood scores. A set of top few optimal estimations is then selected to construct a specific appearance model.

This approach has the implicit assumption that the likelihood of a generic detection will be an effective indicator of its accuracy. Selecting the best few optimal estimations thus will provide accurate detections, which are critical in building a good specific appearance model. However, each optimal estimation is extracted from a different frame. Whilst the frames are all from the same video sequence, they of course will differ in their content. In other words, each optimal estimation is made from different source data (different frames). Hence it is an open question as to whether comparing the likelihood of detections from different frames will actually provide an indication of the relative accuracy of these detections.

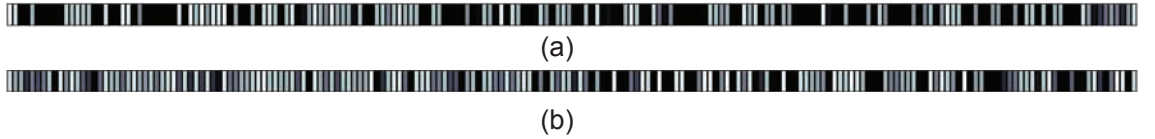


Figure 3.2: An analysis of the correlation between detection accuracy and the likelihood score of the optimal estimations in different frames is conducted, which is based on the Baseball video sequence from [Ramanan *et al.* \(2007\)](#). Each bar represents the likelihood of the optimal estimation in a single frame, sorted on likelihood with intensity indicating ground-truth accuracy. (a) right lower arm. (b) left lower arm. Note that likelihood (order) and accuracy (intensity) have very little correlation.

To this end, an analysis of the Baseball sequence was performed in order to explore the relationship between likelihood and accuracy between frames. The optimal generic detection of each frame was extracted and both its likelihood and accuracy were calculated. Accuracy is in terms of the percentage of overlap between the detection and the ground truth bounding boxes for each body part. Figure 3.2 shows the full set of 200 frames for the optimal estimations of the right lower arm (Figure 3.2-a) and the left lower arm (Figure 3.2-b). These body parts tend to be most difficult to detect accurately and so demonstrate most clearly the analysis conclusions. The figures show the accuracy of each

frame as a coloured bar, where frames are ordered by decreasing likelihood of the detection and intensity indicates how accurate the detection is.

It can be seen that the most accurate (light intensity) detections are relatively randomly distributed. This indicates that there is little useful correlation between detection accuracy and the likelihood score of estimations from different frames. In particular, the top few most likely right lower arm detections include many that are very poor in accuracy. This has the important consequence that the reliance of [Ferrari *et al.* \(2009\)](#) on choosing the most likely optimal estimations to construct the colour appearance model is difficult to justify, and can lead to poor specific colour models. However, this does not mean that the likelihood in the generic detections is meaningless. Within a single frame it is still significant (since the detections are based on the same data), but *between* frames likelihood is not a reliable indicator of accuracy. Specifically, the likelihood of a limb's detection in one frame cannot be compared to the corresponding limb's detection likelihood in another frame: the two likelihoods have little useful ordering for the purposes of inferring relative correctness.

3.4 Investigating the Generic Detector

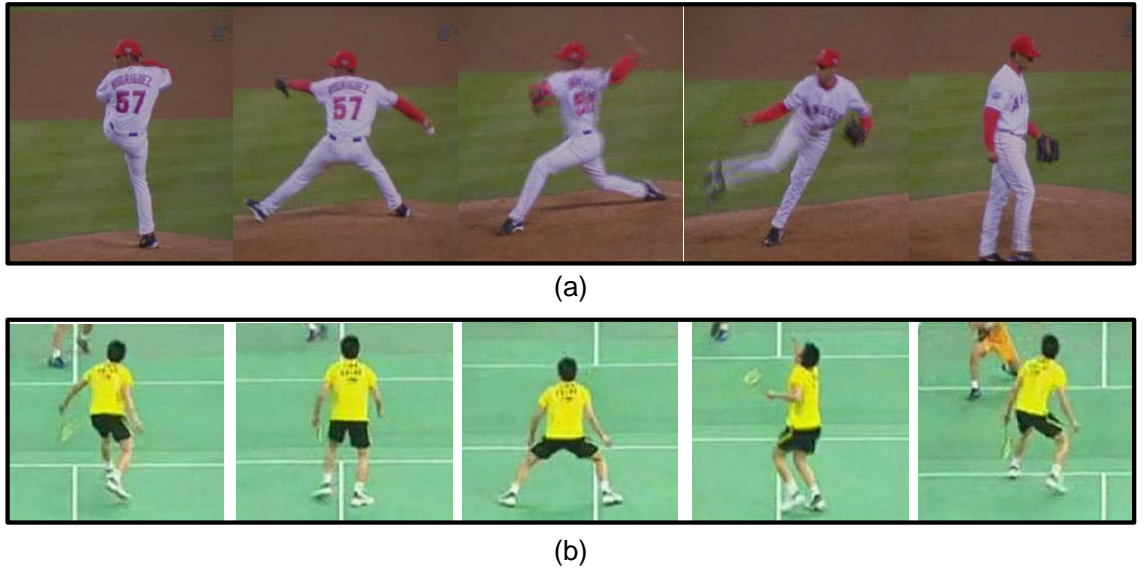


Figure 3.3: (a) The baseball sequence from [Ramanan *et al.* \(2007\)](#). (b) The badminton sequence from online video.

In Section 3.2, the generic detector is built on the framework of the pictorial structure.

When applying the generic detector in estimating the human configuration in each frame of a video sequence, two types of outputs can be obtained by the generic pose detector: one is the optimal estimation for each body part in each frame of the sequence; another is the multiple probable alternative estimations (the sub-optimal estimations) for each body part in each frame of the sequence. This section investigates the generic detector by qualitatively and quantitatively analyzing these two types of outputs. A series of important conclusions are reached. Two sequences, as shown in Figure 3.3, the Baseball sequence used by Ramanan *et al.* (2007) and the Badminton sequence from online video, are used to test this generic detector. The Baseball video is a sequence of 200 frames that records a pitcher throwing out a ball. The Badminton sequence is a sequence of 166 frames that records the motion of a badminton player in a match where the lateral walking pose does not exist. Two videos used in the experiment are representative. First, both baseball video and badminton video record a rich variety of human motion, which meets our requirement for tracked videos, no restriction on the human activities. Moreover, one is captured outdoor and another is captured indoor. Finally, the players in two videos wear normal and different clothes. Therefore, the conclusion reached based on the two videos can be generalized and applied in other videos.

3.4.1 Analysis of the Optimal Estimations

Some examples of the optimal estimations for the Badminton sequence are shown in Figure 3.4. In each image, the optimal estimations for human configuration are shown by red rectangles. Figure 3.4 (a) gives three examples where the optimal estimations are correct comparing with the ground-truth. When a body part has a clear boundary such as those shown in these three examples, it is easily detectable by the generic local detector and result in a high score thus representing a high potential for the human configuration to be correctly estimated. When a body part does not have a clear boundary or is invisible due to self-occlusion, the generic detector is likely to fail in correctly estimating its configuration, such as the estimations for the left upper and lower arms shown in Figure 3.4 (b). It is due to the fact that generic detectors work on the basis of a generic appearance feature (the shape feature in this particular generic detector) which easily results in ambiguity. Especially when the body part is invisible or does not have a clear boundary, a false estimation is much more likely to be chosen as a correct estimation. Furthermore, since the generic detector is easily confused by background clutter, even when there is no self-occlusion or unclear boundary, sometimes false estimations are possibly rated higher than the correct estimations. Some obvious errors could be produced, such as the torso estimations in Figure 3.4 (c). Furthermore, in this thesis, the optimal estimation for each body part is the maximum marginal posterior estimation rather than the joint posterior

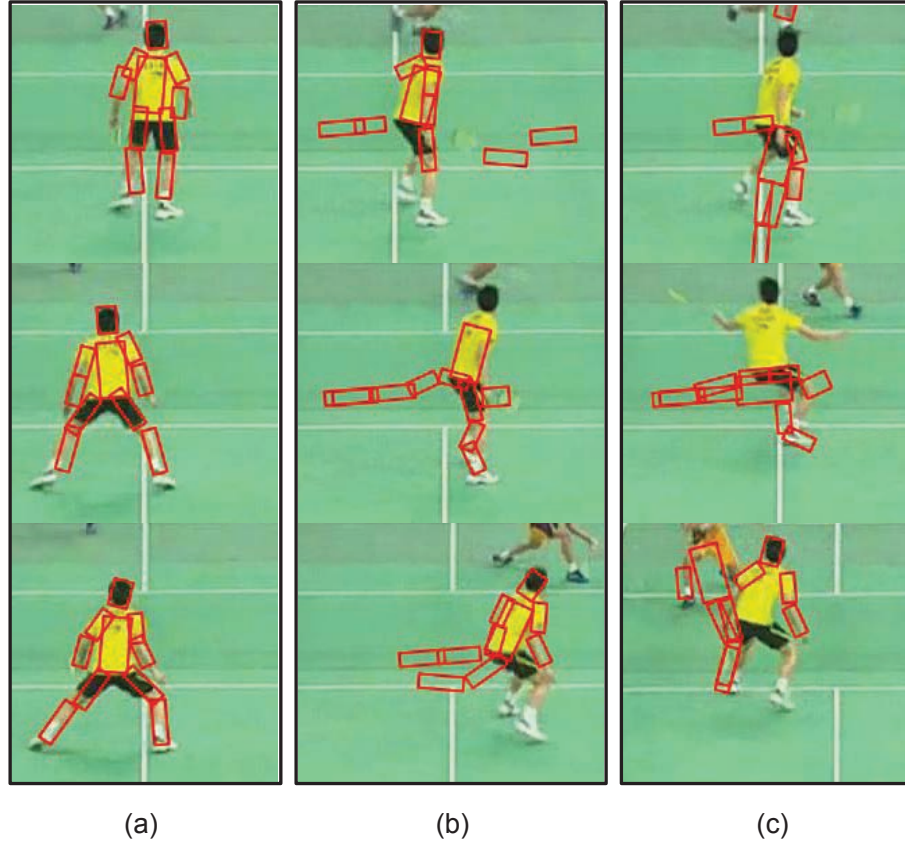


Figure 3.4: Several examples demonstrate that both the correct and incorrect estimations possibly exists within the optimal estimations

estimation. As a result, estimations of two connected body parts such as the torso and the left/right upper arm could be disconnected, such as shown in Figure 3.4

Figure 3.5 provides some examples of the optimal estimations for the Baseball sequence. As with Figure 3.4 (a), three correct estimations for the Baseball sequence is given in Figure 3.5 (a), which further demonstrate that the generic detector works well when body parts are not occluded by each other and there is a clear boundary for each body part. Figure 3.5 (b) gives some other typical self-occlusion examples where the generic detector fails in detecting the lower arms when they are occluded by the torso. The generic detector also tends to fail in detecting human configuration with rare poses, as shown in Figure 3.5 (c). The generic detector considers both the appearance likelihood of all body parts and the spatial relations between them. Even when the correct estimations rate higher in the appearance likelihood than the incorrect estimations, the incorrect estimations are still likely to be chosen as the optimal estimations if they happen to have more suitable spatial relations with other body parts than the correct estimations.

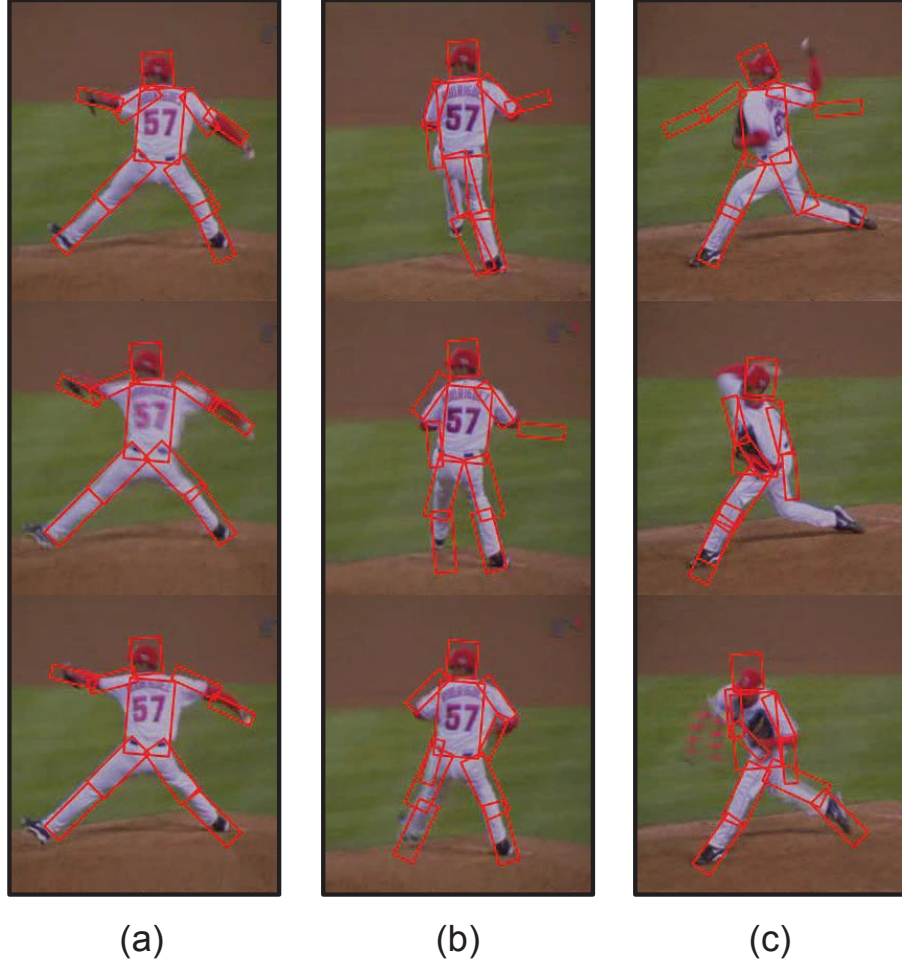


Figure 3.5: Several examples demonstrate that both the correct and incorrect estimations are possibly chosen as the optimal estimations

From the analysis above, it becomes clear that a generic detector tends to fail in the event of self-occlusion, plausible candidates resulting from background clutter, and rare poses. This is caused by the nature of the generic detector. The generic detector only characterizes the generic features rather than the specific features of a human body, thus resulting in its failings in identifying self-occlusion and eliminating plausible candidates. Therefore, it becomes apparent that a specific appearance model for a tracked human (such as the colour appearance model) should be employed to overcome these problems. In addition, a specific appearance model could better track rare poses, since it can easily help to eliminate the plausible candidates.

A quantitative analysis is performed for the correct rate of the optimal estimations in each body part on both the Baseball and Badminton sequences. As shown in Figure 3.6, the optimal estimations for head and torso achieve an high correct rate (over 80%), whereas a

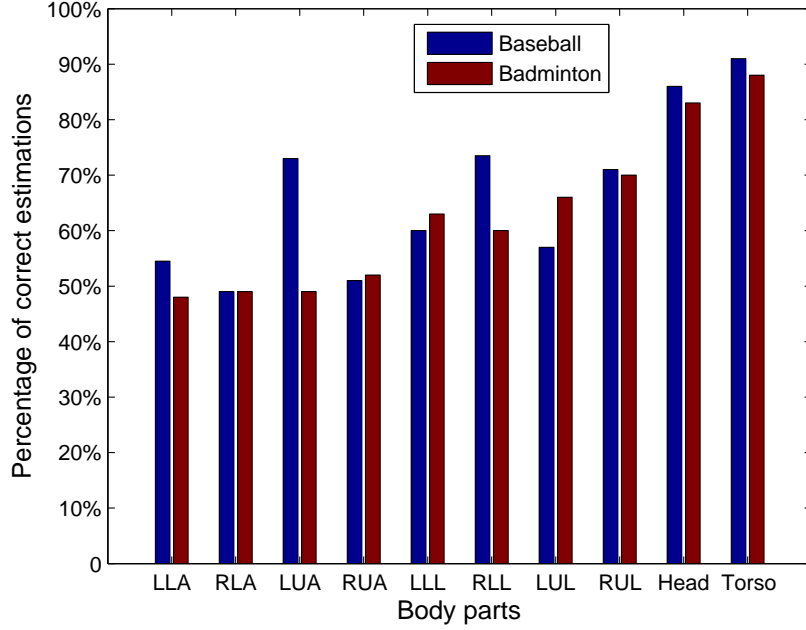


Figure 3.6: The percentage of frames where the optimal estimations are correct for each body part on Baseball and Badminton sequences.

relatively low correct rate is obtained in some other body parts, such as LLA, RLA, LUA and RUA in the Badminton sequence and LLA, RLA, RUA in the Baseball sequence. This poses a problem for further refinement of the optimal estimations, such as what Ferrari did in building a colour appearance model from them. Fortunately, generally at least 50% of the optimal estimations for each body part are correct in the ground-truth sense. This makes it possible for the optimal estimations to be refined using a clustering algorithm, which is proposed in Chapter 4.

3.4.2 Multiple Probable Alternative Estimations

Besides the optimal estimation for each part, a set of sub-optimal estimations for part m , denoted as U_m , is obtained by sampling the top N marginal posterior $P(l_m|I, C)$. We hypothesize that the correct estimation are very likely to lie in this set when the optimal estimation is incorrect. Several experiments are conducted to test and prove this conjecture. The goal is to show that the correct estimate does consistently exist within the top few estimations by the generic detector. This will allow us to develop a method for identifying the correct estimation with alternative features, specifically the colour model described in Chapter 4.

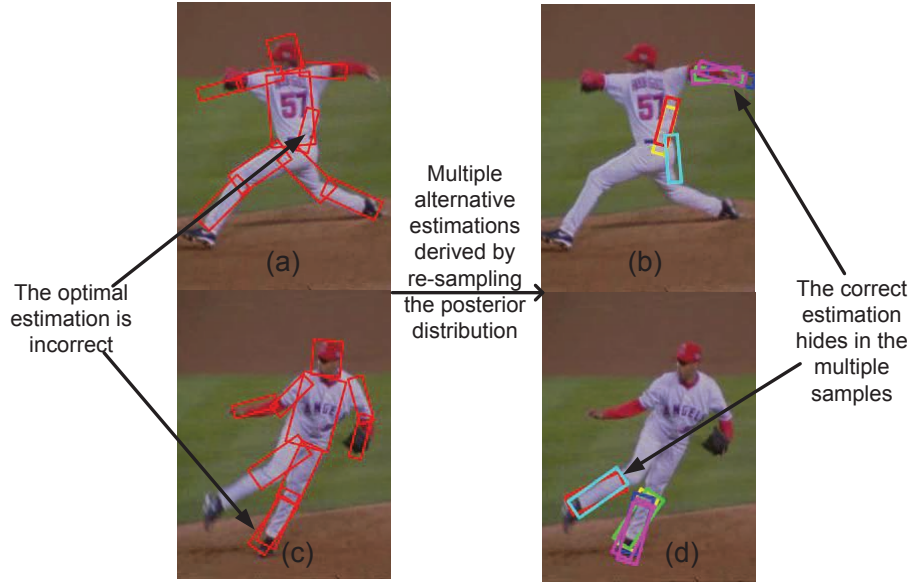


Figure 3.7: Two examples demonstrating why multiple probable alternative estimations are needed. When the optimal estimations are incorrect as shown in Figure (a) and (c), the correct estimations are highly likely to exist in the set of multiple probable alternative estimations as shown in Figure (b) and (d).

Figure 3.7 (a) shows an example where the optimal estimation for the right upper arm (RUA) is incorrect. The multiple probable alternative estimations for RUA, which are the six maximum marginal posterior estimations except the optimal estimation, are sampled and shown in Figure (b). It can be seen that the correct one exists within the multiple probable alternative estimations. A similar case for the left lower leg is shown in Figure 3.7 (c) and (d).

An experiment is conducted to explore whether the correct limb estimation consistently exists within the set of multiple good estimations when the optimal estimation is incorrect and how many samples are needed to ensure that the correct limb estimation appears in the set of multiple good estimations. Figure 3.8 and Figure 3.9 show the percentage of frames where the correct limb estimation exists within the sampled estimations when varying the amount of samples respectively for the Baseball and Badminton sequences. When the number of samples is equal to 1, the set of the sampled estimations only consists of the optimal estimation and the average percentage correctly detected for all body parts is approximately 70% for the Baseball sequence and only about 60% for the Badminton sequence. The percentage of frames where the correct (ground-truth) limb estimations exist within the sampled estimations rises as the number of samples increases and the average percentage reaches about 90% when the number of samples is over 15 for the Baseball sequence and over 20 for the Badminton sequence, which shows that for most

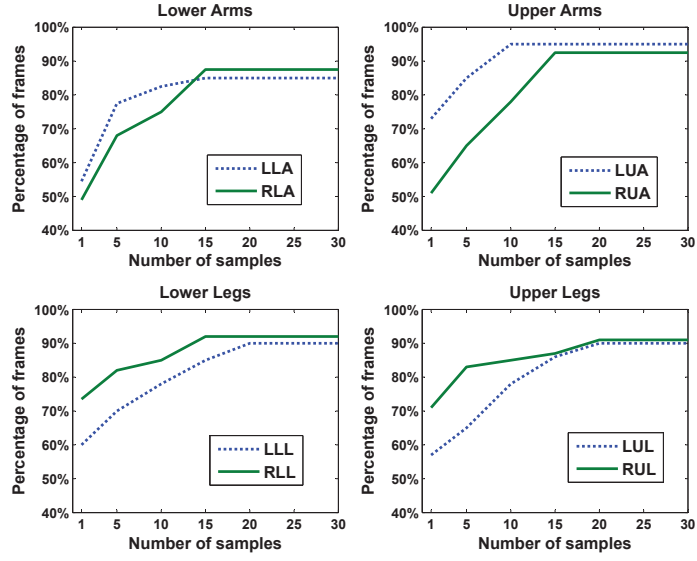


Figure 3.8: Percentage of frames where the correct (ground-truth) limb estimations exist within the sampled estimations. Note that the percentage of frames no longer increases after the number of samples reaches 15. This is because that the correct limb estimations are impossibility to be detected when they are invisible in some cases such as severe motion-blur and self-occlusion. In such cases, the percentage of correct estimation is unable to reach 100% no matter how many possible sampled estimations are checked.

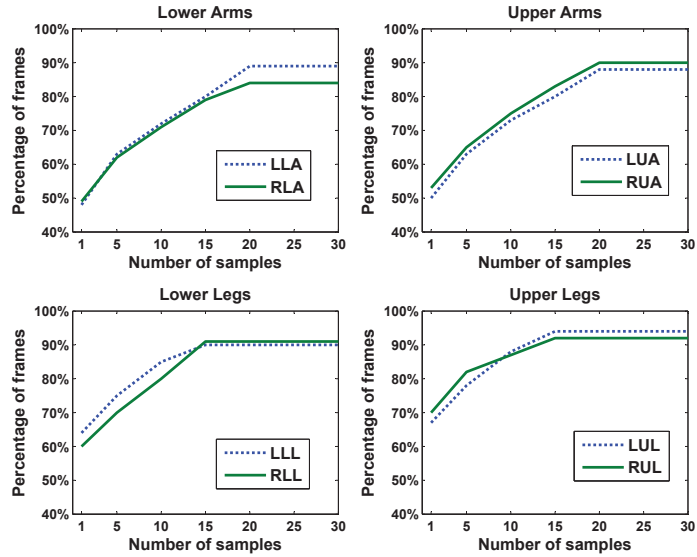


Figure 3.9: Percentage of frames where the correct (ground-truth) limb estimations exist within the sampled estimations on Badminton sequence.

frames if the optimal estimation is incorrect, the correct estimation does exist in the set of multiple probable alternative estimations. Therefore, it is highly likely to find the correct estimation from this set. Note that this percentage no longer increases after a certain number of samples are utilized (15 for Baseball and 20 for Badminton). This is due to the fact that the correct estimations for all body parts are not always possible. They could be invisible due to reasons such as severe motion-blur and/or self-occlusion. In such cases, even a large number of possible estimations are checked, it is possible that none of them are correct. Thus the percentage of frames where the correct estimation exists within the sampled estimations is unlikely reach 100% due to the existence of such cases.

After the analysis above, it can be concluded that the correct estimate does consistently exist within the top few generic detections. Such observation allows us to propose a novel approach (described in Chapter 4) to combine both the generic and specific appearance models for estimating human configurations. Unlike Ferrari’s method where the human configuration estimation is obtained by applying the pictorial structure to the average likelihood between the generic and specific appearance models, a specific appearance model is proposed to filter the multiple probable alternative estimations. Human configurations are detected that are consistent with both the generic and specific appearance models.

3.5 Chapter Summary

This chapter first reviews the framework of the pictorial structure and some related algorithms. Then a generic human pose detector is built based on the pictorial structure. Two types of estimation results are derived from this generic detector: the optimal estimations and the multiple probable alternative estimations (sub-optimal estimations). In order to verify the assumption of Ferrari *et al.* (2009), the relationship between the accuracy and the likelihood of the optimal estimations between different frames were explored. The results show that there is no useful relationship between accuracy and likelihood for comparing optimal estimations between frames. Subsequently, the generic human pose detector is evaluated and analyzed by testing against two representative sequences. Results show that the average correct rate of optimal estimations is between 50%-70% for each body part except the torso and head (whose detection rate is over 80%). However in approximately 90% of frames the correct estimation exists within the multiple probable alternative estimations. Therefore, if a more accurate appearance model for body parts can be learned, it is expected that a correct estimation can be found in the multiple probable alternative estimations based on this accurate appearance model, thus improving the performance of human pose estimations.

In the next chapter, an approach is proposed to learn an accurate appearance model and a system of human pose estimations is built based on this learned accurate appearance model.

Chapter 4

Combining Generic and Specific Appearance Model

In the field of estimating 2D human configurations in a monocular view, bottom-up approaches based on learning a specific (colour-based) appearance model of the tracked person have achieved good performance. Two recent approaches (Ramanan *et al.*, 2007; Ferrari *et al.*, 2009), which have been reviewed and discussed in Chapters 2 and 3, automatically perform such person-specific appearance learning by firstly utilizing a generic detector to roughly estimate the human’s posture in a video and building from these estimates a specific appearance model (usually based on the colour of each limb). This specific appearance model is then applied to the entire video to produce a distribution (probability map) of pose estimations at each frame, either by disregarding the original generic detections (Ramanan *et al.*, 2007) or by averaging the specific and generic detection maps together (Ferrari *et al.*, 2009). From this map the most likely pose is then obtained as the correct pose for the frame.

However, these approaches suffer from some issues resulting from the limitation of either the specific appearance model itself or the learning approach for building the specific appearance model. Incorrect postures are often detected based on a specific appearance model due to the similarity of clothing on different limbs, e.g.: a person’s shirt may be of one colour, hence confusion between the arms and torso can occur when the arms are tucked in or occluded. One approach often used to reduce this is to use temporal smoothing (Ramanan *et al.*, 2007; Ferrari *et al.*, 2009), but it is possible that *incorrect* poses will be propagated. In addition, such approaches usually ignore some useful (non-colour) information on postures provided by the generic detections themselves, which can be utilized to offset the limitation of the specific appearance model. Such information is typically discarded after the colour appearance model is built. Even though Ferrari *et al.* (2009) utilizes the generic detections, they are combined via a weighted linear combination with the specific detections’ probability map. This kind of combination can result in a limb detection to be inconsistent with one of the features, for example, a very strong colour response combining with a weak edge response at the same location can still produce a reasonable score. As for the limitation of the learning approaches for building the

specific appearance model, they either require a specified pose (such as the lateral walking pose in [Ramanan *et al.* \(2007\)](#)) appearing in a sequence or learn a specific appearance by selecting the top few optimal estimations from different frames by comparing their likelihood in [Ferrari *et al.* \(2009\)](#). However, the analysis in Chapter 3 has shown that there is little evidence that the top few optimal estimations from different frames will be accurate estimations.

This chapter proposes an approach for automatically learning a specific appearance model without requiring any pre-specified pose and using a more reasonable assumption than [Ferrari *et al.* \(2009\)](#)'s approach. A tracker is built based on this learned appearance model to produce a final pose estimation that is consistent with both the generic and specific appearance models, and avoids naïvely choosing the colour-based maximum likelihood postures as a result. We initially estimate rough poses using an edge-based generic (shape-focussed) limb detector, which is described in Chapter 3, then cluster these estimations based on their *colour*, with the largest cluster (per-limb) indicating the limb's specific appearance model and other clusters representing background detections. This specific model is then used to filter the generic detections to choose a set of postures that are consistent with the specific appearance. In this way, the specific model is used to verify the generic models rather than simply replace them or average with them. This approach has the benefit of selecting detections that match both the colour and the shape of the limb. In addition, the specific appearance model is a more accurate reflection of the true colour appearance by virtue of clustering, allowing us to effectively separate incorrect detections from correct ones. The final posture is then selected based on its consistency with other body parts. No prior training is needed aside from the generic limb detectors, hence the system can be applied to any video sequence without re-training, unlike the stylized pose detector of [Ramanan *et al.* \(2007\)](#).

Experiments are conducted to compare the proposed system against the approaches of [Ramanan *et al.* \(2007\)](#) and [Ferrari *et al.* \(2009\)](#). Three video sequences are used as the input to compare these systems and the experimental results show that our system is more robust and effective on tracking human in a variety of videos. In addition, our system requires no re-training step to handle a priori unknown motions, unlike [Ramanan *et al.* \(2007\)](#) which requires a particular pose to exist in the video in order to function.

The rest of the chapter is organized as follows: Section 4.1 provides an overview of the proposed system. Section 4.2 introduces how to automatically build a specific appearance model based on mean shift and the preliminary estimations from Chapter 3. Section 4.3 presents the implementation of human pose tracker based on the learned specific appearance model. Section 4.4 presents the experimental results.

4.1 System Overview

In this chapter, we propose a novel algorithm for tracking people by using both the generic and the specific appearance models. Our approach employs a strong (specific) colour appearance model extracted for the tracked target using a generic detector. To avoid the assumption of Ferrari *et al.* (2009) that the most likely generic detections are always correct, clustering is used to avoid contaminating the specific appearance model with incorrect generic detector maxima. Data association is then achieved by using the specific appearance model to filter the generic detections, searching for detections that are consistent with *both* the colour appearance and shape appearance. The proposed approach is implemented in three stages, as shown in Figure 4.1.

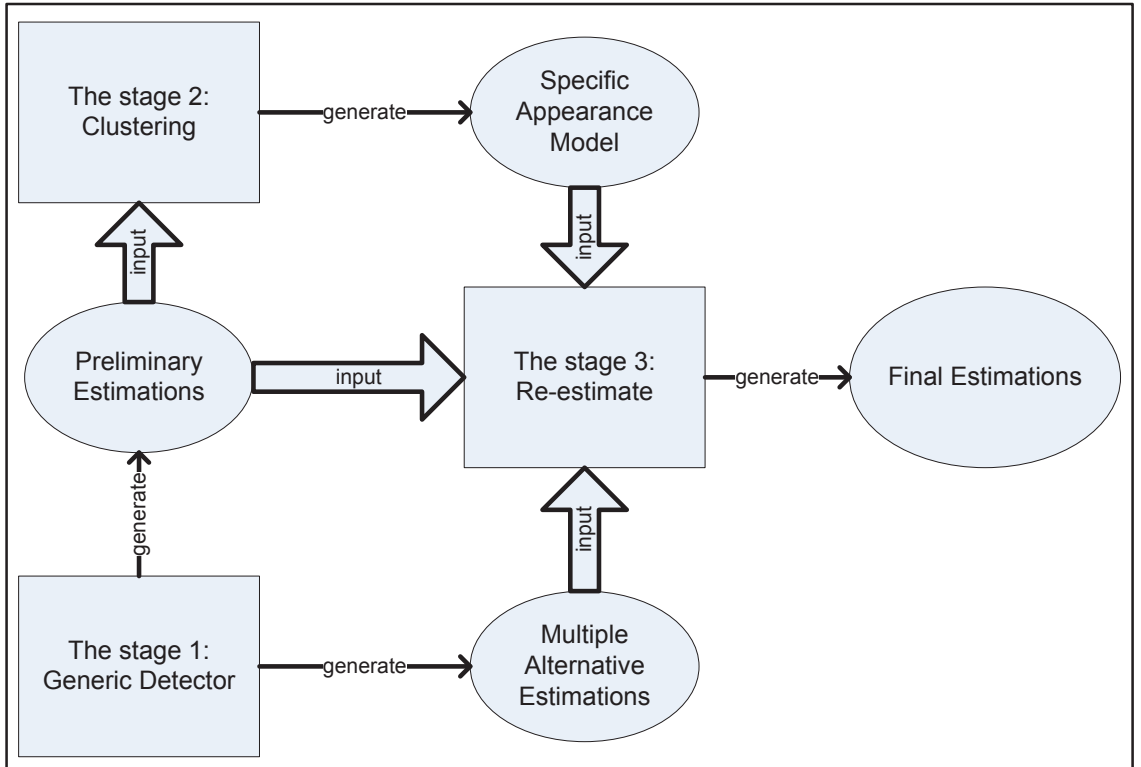


Figure 4.1: The overview of our approach

In Stage 1, the frames are detected using the generic human pose detector described in Chapter 3. In this stage, the optimal estimations with the maximum posterior probability will be chosen as the *preliminary* estimations which will be used in Stage 2 to build a specific appearance model via clustering. However, unlike Ferrari *et al.* (2009), these estimations are not the only configurations considered. A set (M) of multiple configura-

tion estimations for each body part can be generated by sampling the marginal posterior $P(L|I, C)$ which will be used in Stage 3 to identify and improve upon the preliminary estimations that are incorrect given the learned colour appearance.

In Stage 2, the preliminary estimations from Stage 1 are clustered in the colour appearance space to build a specific colour appearance model. Since the colours of all body parts are normally consistent between images in a single video sequence, the correct preliminary estimations of a body part are expected to be gathered to the same cluster representing the correct colour of the body part. In contrast, false preliminary estimations will have a varied colour appearance (depending on what background they overlay) and would thus be dispersed into smaller different clusters, depending on their individual colours. It is expected that the true estimations will be gathered in the largest cluster and the false estimations will be scattered to many small clusters hence the largest cluster is considered to contain the correct appearance estimations. Based on this largest cluster, the specific appearance model is built and used to re-estimate the false preliminary estimations in Stage 3.

In Stage 3, the false preliminary estimations are re-estimated. The estimations in the set M which contains multiple configuration estimations for each human part are checked against the specific appearance model. Since the set M is obtained by re-sampling the marginal posterior, the estimations in set M are firstly constrained by the spatial relations between the body parts and the generic appearance model in the pictorial structure model. The final estimation chosen from the set M is then verified by the specific appearance model. Therefore, the final estimations generated are constrained both by the generic appearance model and the strong appearance model, hence are expected to provide an improved estimate over either of them.

4.2 Building a Specific Appearance Model

In the previous chapter, a generic pose detector is built based on the pictorial structure and edge features. It is applied to estimate the human configuration in each frame of a sequence. Experiments show that the optimal estimations from the generic pose detector are not accurate enough when compared with the ground truth. In this chapter, we seek to find some approach to improve the accuracy of tracking based on the generic estimations. Specifically, we consider these estimations to be a set of noisy pose detections that we must split into ‘accurate’ and ‘inaccurate’ sets. Essentially, the generic optimal estimations are known to be not always reliable. Therefore, the optimal estimations in Chapter 3 will

henceforth be referred to as the preliminary estimations. The analytical results in Chapter 3 also show that the majority of the preliminary estimations are accurate estimations in the sense of the ground truth. Since the two sequences used in the analysis in Chapter 3 are fairly typical and representative of the generic sequences for tracking, it is reasonable to assume that the majority of the preliminary estimations are accurate when applying the generic pose detector in any sequence. Moreover, since the accurate estimations accurately overlay the object being tracked (with some variety of illumination), it is also reasonable to assume that the set of accurate estimations for each body part have similar colour due to the continuity of human appearance in a sequence. In contrast, the inaccurate estimations will represent various colours depending on the errors in the configuration, and these colours will vary from estimation to estimation. For example, as shown in Figure 4.2, the accurate torso estimations from the preliminary estimations have similar colour, whereas the inaccurate torso estimations from the preliminary estimations will have different colours from the accurate estimations.

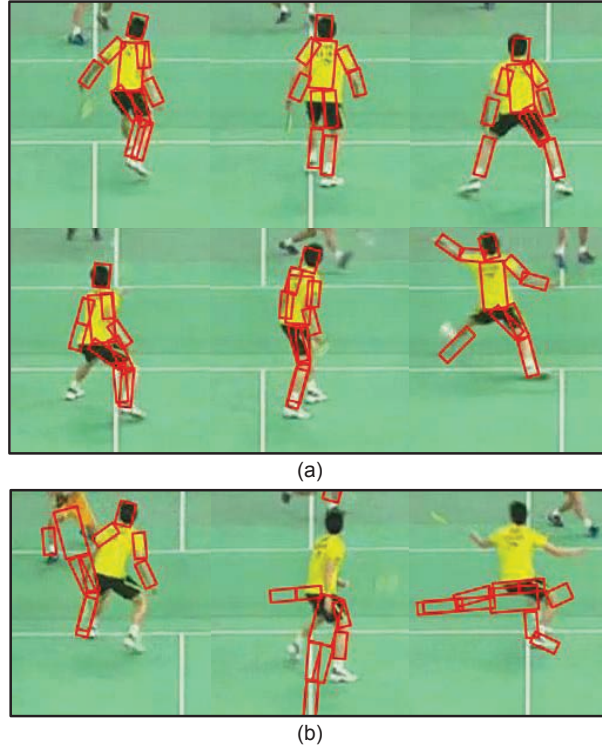


Figure 4.2: (a) The accurate torso estimations from the preliminary estimations have similar colour. (b) The inaccurate torso estimations from the preliminary estimations will have different colours with accurate estimations.

In this section, our aim is to separate the accurate estimations from the inaccurate estimations based on the two assumptions mentioned above. Our main idea is that if the preliminary estimations can be divided into a few clusters using the colour similarity then

the biggest cluster will be the cluster with the accurate estimations. To achieve this aim, the most important issue is to find a colour representation to capture the colour feature of the preliminary estimations. The colour representation is subjected to the requirement that the best-overlapping (accurate) preliminary estimations are close together and far away from the low-overlapping (inaccurate) preliminary estimations in this colour feature space. The clustering algorithm can be applied to make the preliminary estimations with similar colour feature gather together. In this section, the colour representation and clustering algorithm used in our system will be discussed first. An analysis is then conducted using the Baseball video sequence to verify that the chosen feature representation meets the requirement.

4.2.1 Feature Extraction and Representation

Given the preliminary estimations (S_m) for body part m across all frames, the preliminary estimation s_m^t in frame t is denoted by a three-dimensional array $\{x_m^t, y_m^t, \theta_m^t\}$. The three-dimensional array with the predefined width and length of body part m specifies a bounding box where its centre is denoted by $\{x_m^t, y_m^t\}$ and its orientation is denoted by θ_m^t , as shown in Figure 5.9 (a).

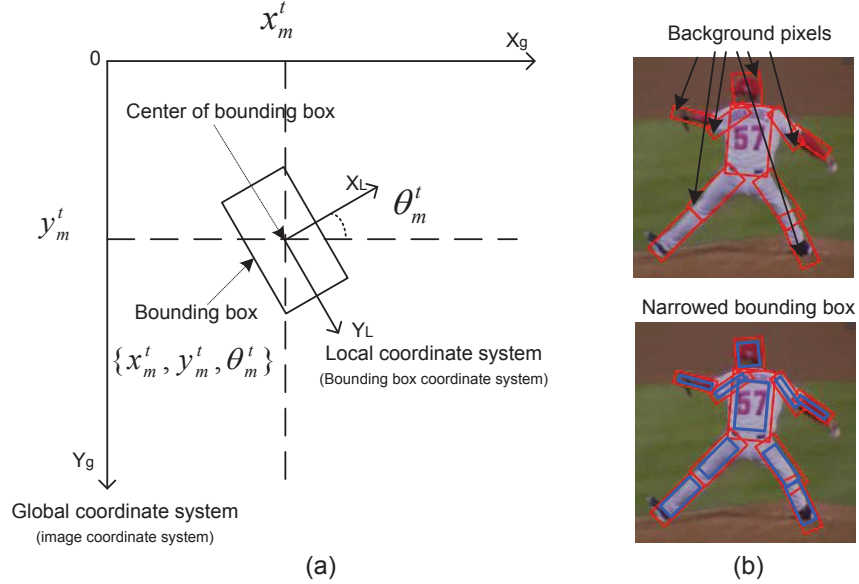


Figure 4.3: A preliminary estimation for body part m in frame t is denoted by $\{x_m^t, y_m^t, \theta_m^t\}$, which corresponds to a bounding box. The center and orientation of the bounding box is respectively denoted by $\{x_m^t, y_m^t\}$ and θ_m^t .

A bounding box is never a perfect fit to a true body part thus some background pixels

are inevitably included even in an accurate preliminary estimation. As shown in Figure 5.9 (b), aside from the pixels representing the tracked object there exist some pixels that represent the environment/background since the bounding box is usually larger than the body part and perhaps not quite aligned with it. When extracting the appearance feature specified by a bounding box, in order to reduce the inclusion of background pixels, the bounding box is narrowed slightly as shown in Figure 5.9 (b) for the purpose of extracting colour features of the corresponding body part. The width and length of the narrowed bounding box are set as 70% of the original bounding box. This proportion gives a good balance between including too much background (for too large a rectangle) or not extracting enough pixels from the limb itself (if the bounding box is too narrow).

A feature vector (v_m^t) is constructed to capture the appearance feature of the preliminary estimation (s_m^t). A normalized colour histogram in the L*a*b* colour space is used to represent the appearance feature of this preliminary estimation. The histogram is represented by separately projecting the pixels inside the corresponding narrowed bounding box onto the L, a, b axes in the L*a*b* colour space. Each of the three colour channels is divided into 10 evenly distributed bins. The feature vector thus consists of a 30-dimensional L*a*b* colour histogram. The set of the preliminary estimations (S_m) for part m is then transformed to a set of feature vectors $\{v_m^t\}$ in the L*a*b* histogram feature space. Although there are many possible colour features (eg: different colour space, etc), we find in Section 4.2.3 that L*a*b* histogram feature works well, i.e., meeting our requirement.

4.2.2 Mean Shift for Clustering

After the feature representations for the preliminary estimations are defined, in order to gather the preliminary estimations with similar colour feature, an applicable clustering algorithm must be chosen. There are a huge number of publications in the field of clustering. Some methods are based on the number of clusters in the feature space and some others are based on the shape of clusters in the feature space. Such methods are not applicable here, since both the number of clusters and the shape of clusters in the feature space are unknown in our case. Therefore, mean shift (Comaniciu and Meer, 2002), a nonparametric clustering method, is chosen for our system. To use mean shift to detect clusters, the corresponding space feature is viewed as an empirical probability density function (p.d.f). The procedure of finding clusters is actually the procedure of finding the dense regions in the feature space. The dense regions would correspond to the local maximas of the p.d.f, which are the modes of the distribution. The mean shift procedure is used to find these modes, and consequently find the corresponding clusters.

The set of colour histogram feature vectors $\{v_m^n\}$ for body part m in a sequence can be viewed as N data points (v_m^n) , $n = 1, \dots, N$ in the 30-dimensional space R^{30} described by the histograms. The density estimator for the analyzed feature space is defined as

$$\hat{f}(v) = \frac{c_{k,d}}{N \times h^d} \sum_{t=1}^N k(\|\frac{v - v_m^n}{h}\|^2) \quad (4.1)$$

where $c_{k,d}$ is the normalization constant and k is the following kernel:

$$k(x) = \exp(-\frac{1}{2}x) \quad x \geq 0 \quad (4.2)$$

The mean shift procedure can start from any data point in the feature space and stop in a convergence point where the mode is located. When the constant h is fixed, a data point v will correspond to a unique convergence point.

In our system, the mean shift procedure is starting from the points in the set of feature vectors $\{v_m^n\}$ and a set of convergence points (modes) $\{z_m^n\}$ is obtained where each v_m^n corresponds exclusively to a convergence point (mode) z_m^n . The constant h is determined according to the real size of body part in the image. In the baseball sequence, the size of torso in the frame is about 50 pixels width and 70 pixels length. The h used in our experiment is 330. For other body parts, as the size decreases, the h is reduced correspondingly. Given the length L and width W of a body part, the h for that body part can be calculated using the following equation, $h = \frac{33 \times L \times W}{350}$.

After the convergence point z_m^n for each feature vector v_m^n is obtained, clustering can be made by computing the distance between the convergence points. For any two feature vectors, if their convergence points are close enough, they can be put into the same cluster. Formally, the clusters $\{\beta_p\}$ are then derived if the following conditions are satisfied:

1. $\forall v_m^a, v_m^b \in \beta_p, a \neq b, \|z_m^a - z_m^b\| \leq h_s$, where z_m^a, z_m^b correspond respectively to v_m^a, v_m^b and h_s is a parameter constant.
2. The clusters $\{\beta_p\}, p = 1 \dots s$ are obtained which satisfy:

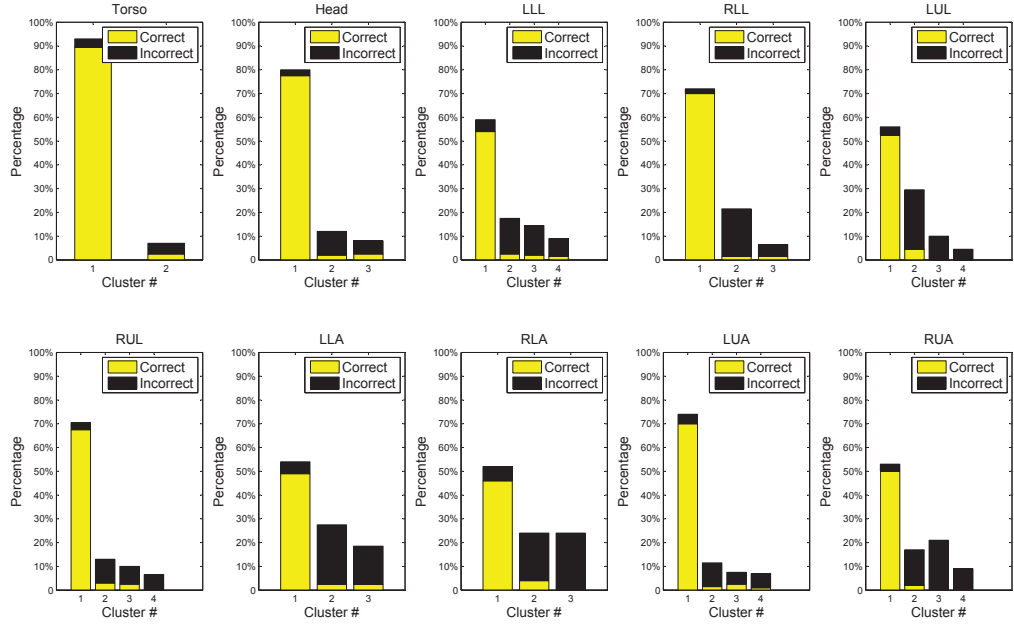
$$\cup_{p=1}^s \beta_p = \{v_m^n\}, \quad (4.3)$$

$$\cap_{p=1}^s \beta_p = \phi. \quad (4.4)$$

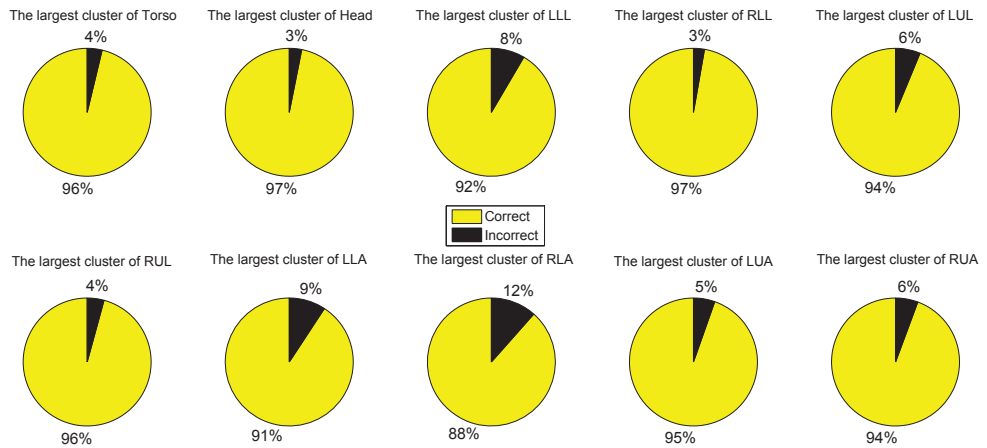
Normally, the parameter constant h_s is set as a very small value. In our experiment, the h_s is set as 2. The parameter constant h_s can be adjusted to control the merging of the clusters. The clusters representing similar colours can be merged by increasing h_s .

4.2.3 Analysis of the Clusters Obtained by Mean Shift

After the features of the preliminary estimations are captured by colour histogram vectors, a few clusters are obtained using mean shift. According to the two previous assumptions that the accurate estimations have similar colour feature and the majority of preliminary estimations are accurate in the sense of the ground truth, it is important that the preliminary estimations with similar colour feature are gathered together and most of the preliminary estimations in the largest cluster are accurate.



(a) The results of clustering for each body part



(b) The components of the largest cluster for each body part

Figure 4.4: The analysis for the results of clustering.

In order to prove that this supposition is applicable in our approach, an experiment is conducted for the Baseball sequence to analyze the components of the clusters, i.e., to compute how many accurate and inaccurate estimations there are in each cluster. To implement this analysis, the preliminary estimations for each body part are first labelled as ‘accurate’ or ‘inaccurate’ according to the ground truth. Specifically, an estimation is considered an accurate estimation when the overlapping rate between this estimation and its ground truth is over 75%, otherwise it is considered an inaccurate estimation. The components of any cluster can then be analyzed. The results of clustering for each body part are shown in Figure 4.4(a) where each cluster is represented by a bar and the correct and incorrect estimations are respectively represented by the yellow and black parts. It can be seen that the vast majority in the largest cluster are accurate estimations although a small part of this cluster are inaccurate estimations. Figure 4.4(b) provides a more detailed breakdown of the components in the largest cluster. For any body part, the accurate estimations (preliminary estimations) in the largest cluster are between 88% and 97% of all estimations (preliminary estimations) in this cluster. If the estimations in the largest cluster are used as training samples to learn a specific appearance model, even for the worst performing limb of RLA, 88% of these training samples represent the accurate appearance of the respective body part.

In order to automatically build a specific appearance model, Ferrari *et al.* (2009) also proposed an approach to refine the preliminary estimations based on the most likely detections. More specifically, Ferrari *et al.* (2009) ranks the preliminary estimations by their likelihood scores. A set of most likely preliminary estimations is then selected to construct a specific appearance model. However, in Section 3.3, an analysis of the Baseball sequence shows that even the detections in the frames containing the most likely preliminary estimations are not necessarily more accurate than those in the other frames. Hence the reliance of Ferrari *et al.* (2009) on choosing the most likely preliminary estimations to construct the colour appearance model can lead to poor specific colour models. According to the experimental results in Chapter 3, in Baseball sequence, only about 50% of all preliminary estimations for RLA is accurate (see Figure 3.6). If Ferrari *et al.* (2009)’s approach is used to select the training sample of RLA’s specific appearance model, only 50% of the training samples are the accurate estimations. However, in the proposed approach, 88% of training samples are the accurate estimations. This demonstrates that clustering significantly improves the quality of estimations identified to be used subsequently in building a colour model since it discards most of the inaccurate generic detections.

4.2.4 Specific Appearance Model

From the previous analysis, it can be seen that the preliminary estimations can be divided into different clusters based on their different colour features and the largest cluster can represent the correct estimations. In this section, a specific appearance model will be built based on such observations. As discussed in Section 4.2.2, each feature vector in a cluster has a corresponding mode and these modes are close enough in the feature vector space. Based on these modes, an average mode (mean) μ and a variance σ can be computed for each cluster. Thus, the colour feature of each cluster can be represented by a 2D vector $\langle \mu, \sigma \rangle$ consisting of its average mode (mean) and variance. Given a body part m and K clusters of the preliminary estimations for part m across all frames, a set of 2D vectors $\{\langle \mu_m^k, \sigma_m^k \rangle | 1 \leq k \leq K\}$ are used to define the specific appearance model of part m . In this set, there is a vector $\langle \mu_m^*, \sigma_m^* \rangle$ corresponding to the largest cluster and representing the colour feature of the target body part and several other vectors corresponding to other small clusters and representing the non-target part or the background colour feature.

4.3 Tracking

After the specific appearance model is learned in the second stage as described in Section 4.2, we need to utilize this specific appearance model for a second pass that produces more accurate trackings as Ramanan *et al.* (2007) (colour-only) and Ferrari (average of colour and edges) have done. As discussed in Chapter 2, Ramanan *et al.* (2007) builds a specific (colour) appearance model using the results of a ‘stylized pose’ detector that can only detect a specified pose, i.e., the lateral walking pose. The learned specific appearance model is then applied in a pictorial structure for a second pass, i.e., re-estimate human pose based on the learned specific appearance model in each frame. Since the appearance model used in the pictorial structure is colour-only, some issues will happen during the second pass. The weaknesses of the approaches based on colour-only appearance model (refer to Section 2.5.3) are recapped here.

Firstly, when the upper arms have the same colour with the torso, the colour-only appearance model of the upper arms is easily confused with the colour appearance of the torso thus causing incorrect estimations. These confusions are possibly offset by correctly identifying the corresponding lower arms. However, when the lower arms are occluded by torso, which often happens, inaccurate estimations for the upper arms are almost unavoidable. Secondly, a colour-only appearance model usually leads to inaccuracy in estimating the orientation of human body parts.

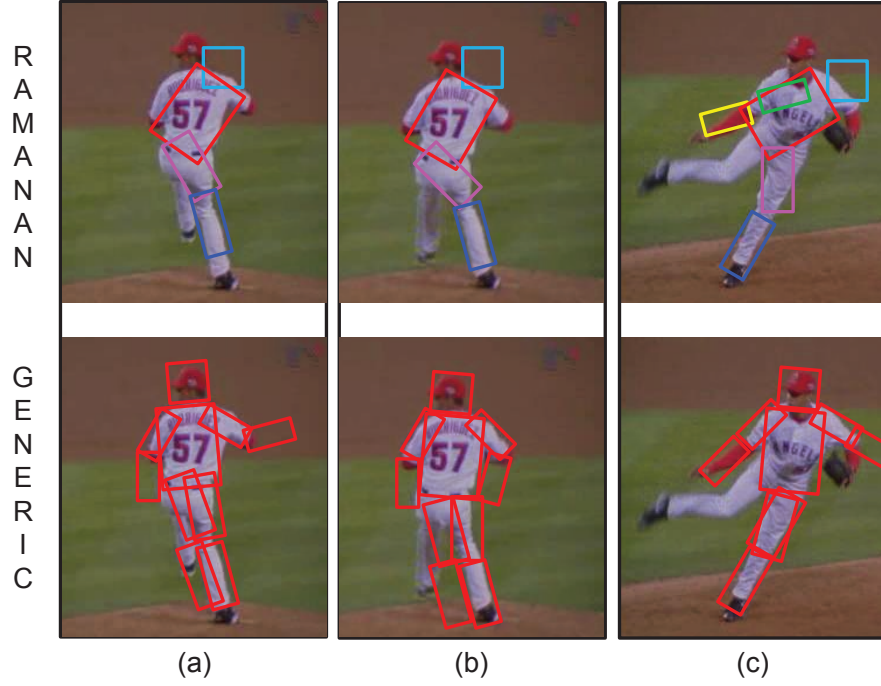


Figure 4.5: Some examples to compare the estimations from [Ramanan *et al.* \(2007\)](#)’s colour-only detector with the estimations from the generic detector described in Chapter 3.

Fortunately, the two issues can be in some extent avoided in a pose detector using a generic (edge-based) appearance model. Figure 4.5 gives some examples to compare the estimations from the generic detector described in Chapter 3 with the estimations from [Ramanan *et al.* \(2007\)](#)’s colour-only detector. When the lower arms are occluded by the torso, [Ramanan *et al.* \(2007\)](#)’s approach cannot estimate the configuration of the upper arms, as shown in Figure 4.5 (a) and (b). Moreover, the estimations for the torso orientation are easily inaccurate. In contrast, the generic detector can obtain accurate estimations for the upper arms even though the lower arms are occluded. In addition, in [Ramanan *et al.* \(2007\)](#)’s approach, if the estimations for the lower arms and torso are available, the estimations for the upper arms are consequently obtained. However, if the estimation for torso orientation is inaccurate, the consequent upper arm estimations are also incorrect, as shown in Figure 4.5 (c).

[Ferrari *et al.* \(2009\)](#)’s approach attempts to combine the generic and specific appearance model to overcome the issues resulting from using a colour-only appearance model, in which the average likelihood of the specific and generic appearance models are applied in the pictorial structure. However, this approach does not fundamentally solve the problems discussed. For example, when the likelihood of the colour appearance model is very high,

even if the likelihood of the generic appearance model is low, the average of the two can still be high. Thus, the issues resulting from the colour-only appearance model will still exist.

Our aim is to find an approach which retains the advantage but overcomes the disadvantage of the colour-only appearance model, thus addressing the issues discussed above directly. Unlike Ferrari *et al.* (2009)’s approach where estimations are based on a likelihood which is the simple average of a colour appearance model and a generic appearance model, our key idea is to obtain estimations that are accepted by both the specific appearance model and the generic appearance model, i.e., to find the estimations that satisfy both the specific appearance model and the generic appearance model. The generic detections from the generic pose detector in Chapter 3 and the specific appearance model based on the clustering method discussed in Section 4.2 provides us with an opportunity to fuse generic detections with specific appearance model in a different way to Ferrari *et al.* (2009)’s averaging method.

4.3.1 Fusing Generic and Specific Appearance Detections

As described in Chapter 3, the generic pose detector generates two types of generic detections, i.e., the preliminary estimation and multiple alternative estimations. In any frame, the preliminary estimation for body part m is the optimal (posterior) estimation, which is the best match with the generic appearance model under the pictorial structure. Beside the preliminary estimation, a set of sub-optimal estimations (i.e., multiple alternative estimations) for body part m , denoted as U_m , is obtained by sampling the top N posterior excluding the maximum posterior. The estimations in this set will also be good matches to the generic appearance model under the pictorial structure. The aim in this section is to find the estimation that satisfies both the generic and specific appearance models. The experiment in Section 3.4.2 has demonstrated that the correct estimation consistently exists within the top few generic detections, providing us with good reason to believe that using a different appearance (colour) model could potentially identify this correct estimation. Therefore, the idea is to search for the correct estimation inside these sampled generic detections (including the preliminary estimation) using the specific appearance model.

This process involves two steps. Firstly, the preliminary estimation is verified using the specific appearance model. If the preliminary estimation matches the specific appearance model, this preliminary estimation will be accepted as the final estimation. In such a case, the final estimation is considered as a highly confident estimation, which will be used as a reference in the process of local search. Secondly, if the preliminary estimation does not

match the specific appearance model, the candidates are obtained by filtering the multiple alternative estimations using the specific appearance model and the final estimation is determined by local search, which is discussed in Section 4.3.2.

4.3.1.1 Verify the Preliminary Estimation

As discussed in Section 4.2, the preliminary estimations for body part m across all frames are divided into several clusters where the largest cluster is assumed to represent the correct estimations. In other words, the preliminary estimations that exist in the largest cluster satisfy the specific appearance model. Therefore, given a preliminary estimation for body part m in frame t , it is easy to verify whether the preliminary estimation satisfies the specific appearance model by confirming if the preliminary estimation is located in the largest cluster. If the preliminary estimation matches the specific appearance model, it is the optimal estimation that satisfies both the generic and specific appearance models under the existing information and thus is accepted as the final estimation for body part m in frame t . The final estimation derived in this way is considered as a highly confident estimation. It will be used as a reference in the process of local search.

4.3.1.2 Filtering Multiple Alternative Estimations

If the preliminary estimation does not match the specific appearance model, the set of sub-optimal estimations (i.e., multiple alternative estimations) U_m are filtered using the specific appearance model. The filtering is the process of removing the elements of U_m that do not agree with the specific appearance model, thus generating a subset U_m^* in which every element satisfies both the generic and specific appearance models. This includes going through estimations that fall within the largest cluster. The subset U_m^* is called the filtered set of sub-optimal estimations and each estimation in U_m^* is a candidate for the final estimation.

As discussed in Section 4.2, the specific appearance model of body part m , is defined as a set of 2D vectors $\{\langle \mu_m^k, \sigma_m^k \rangle | 1 \leq k \leq K\}$ in which the vector $\langle \mu_m^*, \sigma_m^* \rangle$ represents the correct appearance. Given an estimation u in the subset U_m , the filtering based on the specific appearance model is achieved as following:

1. transform the estimation u to an colour histogram vector v' ;

2. find its mode z' using the mean shift algorithm;
3. if $|z' - \mu_m^k| < 2\sigma_m^k$, the estimation u belongs to the k_{st} cluster.
4. if $|z' - \mu_m^*| < 2\sigma_m^*$, the estimation u is verified as matching the specific appearance model.

Figure 4.6 illustrates the process of filtering in a more intuitive way. As shown in Figure 4.6 (a), preliminary estimations for body part m across all frames are divided into several clusters. The largest cluster (also called the main cluster in the figure) represents the estimations with correct appearance and other clusters represent the estimations with incorrect appearance. In Figure 4.6 (b), the elements that belong to multiple alternative estimations U_m are represented by a solid circle. Given an element in the U_m , if this element locates in the main cluster, which means that it satisfies the specific appearance model, it is put into the subset U_m^* . The elements are retained in U_m^* for further analysis in Section 4.3.2 to select the final estimation.

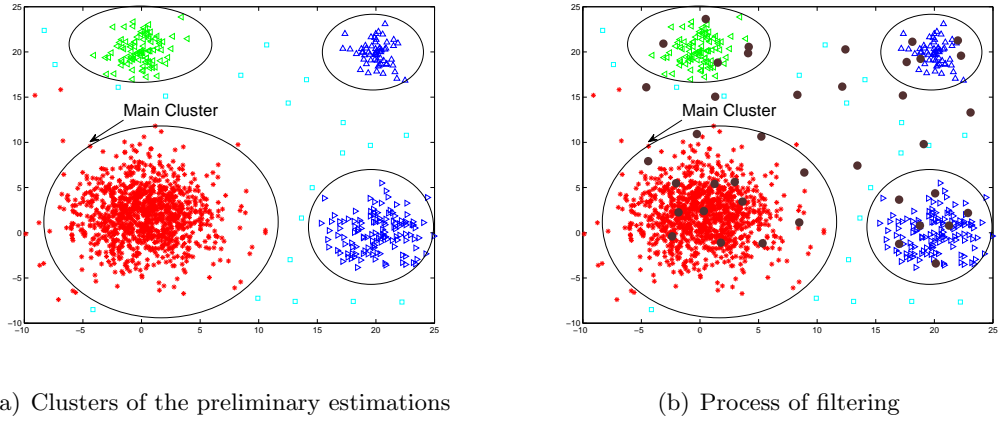


Figure 4.6: Process of filtering. Symbols of solid circle represent the elements in the set U_m and other symbols represent the elements in the set S_m . The elements locating in the main cluster are selected into the set U_m^* .

4.3.2 Local Search

After the set of the filtered sub-optimal estimations U_m^* is obtained, the final step is to determine which candidate in this set is chosen as the final estimation. The most direct approach is to sort the candidates in the set U_m^* according to their corresponding posteriors from the generic pose detector and choose the highest-posterior candidate as the final estimation. However, such an approach ignores an important characteristic of

human motion, specifically, the continuity of human motion in time-space. Instead, the final estimations are determined by considering both the posteriors of candidates and the continuity of human motion. This is implemented by a local search in both time (across frames) and space (between body parts).

The local search can effectively reduce the space of search by applying the continuity of human motion, thus improving the efficiency of algorithm. Meantime, the local search also can slightly improve the accuracy of pose estimation, but it is not essential.

4.3.2.1 Temporal Search

Given the filtered subset U_m^{*t} that contains the candidates for body part m at frame t , the aim of temporal search is to find the optimal estimation from the set U_m^{*t} , in which the continuity of human motion is satisfied. The precondition of the temporal search for body part m in frame t is that the final estimation for body part m in a neighbouring frame (previous or next) is a highly confident estimation (i.e., the final estimation is the preliminary estimation that also satisfies the specific appearance model), which can provide a reliable reference for temporal search. A local temporal search is performed to generate the subset $U_m'^t \subseteq U_m^{*t}$. Moreover, each candidate in the set $U_m'^t$ is required to be in agreement with the location of the final estimation in the previous (or next) frame for the specific body part. In other words, The subset $U_m'^t$ is obtained by filtering the set U_m^{*t} according to the final estimation for body part m in the previous (or next) frame. This filtering process can be expressed formally as follows. Given the final estimation $l_m^{t-1} = \{x, y, \theta\}$ for body part m in frame $t - 1$, any candidate $\{x', y', \theta'\}$ in the set $U_m'^t$ should satisfy:

$$\begin{cases} \{x', y', \theta'\} \in U_m^{*t}, \\ x - p \leq x' \leq x + p, \\ y - q \leq y' \leq y + q, \\ \theta - \alpha \leq \theta' \leq \theta + \alpha, \end{cases}$$

where p, q, α are the parameters that defines the range of temporal search. In our experiments, the values of these parameters can be changed according to different scales. When $scale = 1$, p and q are set as 10, which is the pixel distance. The angle parameter α is set as $\pi/18$. In fact, the setting of parameters depends on the velocity of the human motion in the real world. Since the velocity of human motion changes in a fixed range in the real world, the range of parameters can be determined. Finally, the highest-posterior candidate in the $U_m'^t$ is selected as the final estimation for part m in frame t .

To avoid error propagation, the local temporal search is employed only if the final estimation in the previous (or next) frame is a highly-confident estimation. In contrast, if the final estimations for the previous frame and next frame are obtained by local search (not the highly-confident estimations), the local spatial search is performed to find the final estimation for the current frame.

4.3.2.2 Spatial Search

Before discussing the spatial search, we first define the concept of hinge points for each body part in a human body model and describe how to compute the coordinates of the hinge points for a body part given its configuration.

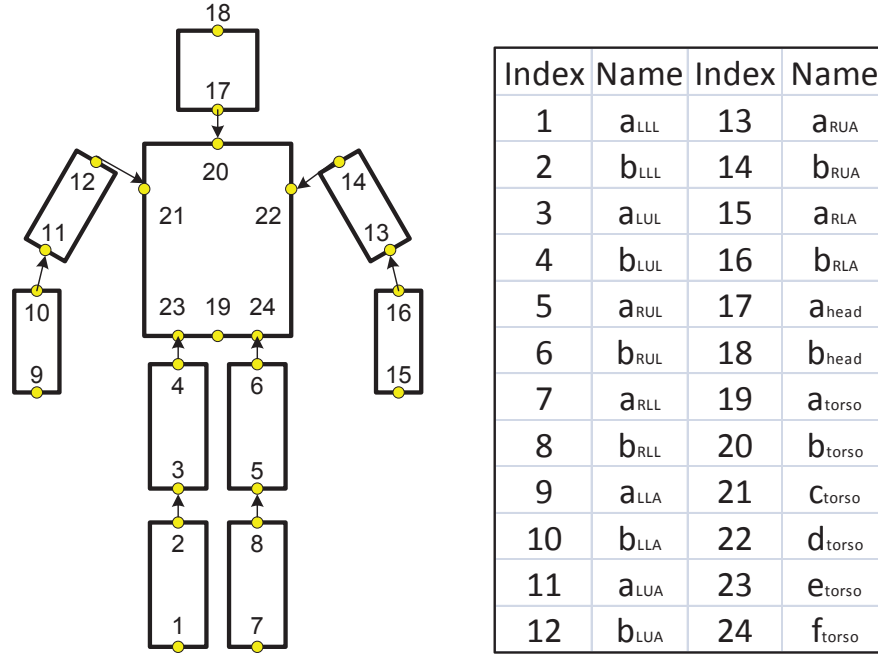


Figure 4.7: The endpoints for all body parts in the human model.

As shown in Figure 4.7, 24 hinge points are defined for our human model. The torso has five hinge points, which are respectively denoted as a_{torso} , b_{torso} , c_{torso} , d_{torso} and e_{torso} . All other body parts have two hinge points, which are denoted as $a_{nameofpart}$ (or a_m) and $b_{nameofpart}$ (or b_m). Any two connected body parts are connected by a pair of hinge points — one hinge point from a child part is dependent on another hinge point from its parent part. For example, the hinge point b_{LUA} (point 12) is dependent on the hinge point c_{torso} (point 21).

Given the configuration $l_n^t = \{x_n^t, y_n^t, \theta_n^t\}$ for body part n in frame t and its length L and width W , the coordinate of the hinge point a_n^t (denoted as $\{x_a, y_a\}$) and the coordinate of the hinge point b_n^t (denoted as $\{x_b, y_b\}$) in the image coordinate system can be computed using the following equations:

$$\begin{aligned} \{x_a, y_a\} &= \{x_n^t + \frac{L}{2} \sin \theta_n^t, y_n^t + \frac{L}{2} \cos \theta_n^t\}, \\ \{x_b, y_b\} &= \{x_n^t - \frac{L}{2} \sin \theta_n^t, y_n^t - \frac{L}{2} \cos \theta_n^t\}. \end{aligned} \quad (4.5)$$

If body part n is the torso, there are another four hinge points c_n^t, d_n^t, e_n^t and f_n^t . Their coordinates are respectively denoted as $\{x_c, y_c\}, \{x_d, y_d\}, \{x_e, y_e\}$ and $\{x_f, y_f\}$, which can be computed formulas as follows:

$$\begin{aligned} \{x_c, y_c\} &= \{x_n^t - \frac{W}{2} \cos \theta_n^t - \frac{L}{4} \sin \theta_n^t, y_n^t + \frac{W}{2} \sin \theta_n^t - \frac{L}{4} \cos \theta_n^t\}, \\ \{x_d, y_d\} &= \{x_n^t + \frac{W}{2} \cos \theta_n^t - \frac{L}{4} \sin \theta_n^t, y_n^t - \frac{W}{2} \sin \theta_n^t - \frac{L}{4} \cos \theta_n^t\}, \\ \{x_e, y_e\} &= \{x_n^t - \frac{W}{4} \cos \theta_n^t + \frac{L}{2} \sin \theta_n^t, y_n^t + \frac{W}{4} \sin \theta_n^t + \frac{L}{2} \cos \theta_n^t\}, \\ \{x_f, y_f\} &= \{x_n^t + \frac{W}{4} \cos \theta_n^t + \frac{L}{2} \sin \theta_n^t, y_n^t - \frac{W}{4} \sin \theta_n^t + \frac{L}{2} \cos \theta_n^t\}. \end{aligned} \quad (4.6)$$

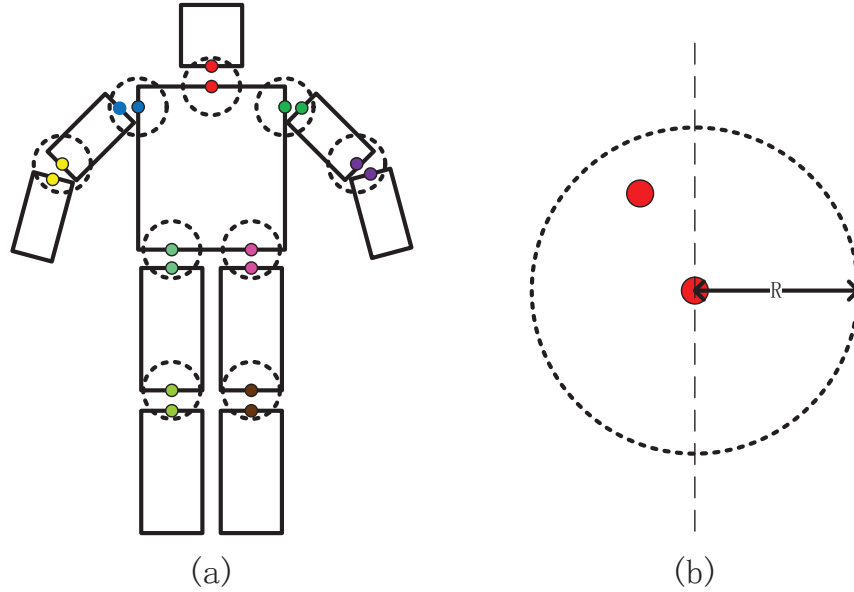


Figure 4.8: The endpoints for all body parts in the human model.

The aim of the spatial search is to find an estimation from the set U_m^{*t} that fits with the neighbouring body parts. Moreover, this estimation for body part m is required to satisfy the spatial constraint of its parent body part n . The spatial constraint applied to body part m (child) from body part n (parent) is described as follows. As shown in

Figure 4.8, ten human body parts are connected by nine pairs of hinge points. For any pair of hinge points, the hinge point of the child part is required to exist within the circle of radius R centred at the hinge point of the parent part. Formally, for a pair of hinge points, given the coordinate of the hinge point $\{x_{h_n}, y_{h_n}\}$ in the parent (body part n), the coordinate of the hinge point $\{x_{h_m}, y_{h_m}\}$ in the child (body part m) is required to satisfy

$$(x_{h_m} - x_{h_n})^2 + (y_{h_m} - y_{h_n})^2 < R. \quad (4.7)$$

Given the filtered set U_m^{*t} that contains the candidates for body part m at frame t , the subset of U_m^{*t} , denoted as $U_m'^t$, which contains the candidates of body part m satisfying the spatial constraint from its parent (body part n), are obtained by further filtering the set U_m^{*t} . Specifically, given the final estimation $l_n^t = \{x_n^t, y_n^t, \theta_n^t\}$ for the parent (body part n) in frame t , the coordinate $\{x_{h_n}, y_{h_n}\}$ of corresponding hinge point is computed using Equation (4.5) and (4.6). For any candidate $l_m^t = \{x_m^t, y_m^t, \theta_m^t\}$ in set U_m^{*t} of body part m , if the coordinate $\{x_{h_m}, y_{h_m}\}$ of its hinge point satisfies Equation (4.7), the candidate is put into the subset $U_m'^t$ otherwise discarded. Finally, the highest-posterior candidate in subset $U_m'^t$ is chosen as the final estimation.

4.4 Experiments and Discussion

The proposed system is tested on several video sequences to evaluate its performance. It is compared against the approaches of Ramanan *et al.* (2007) based on their provided source code as well as Ferrari *et al.* (2009). However, Ferrari *et al.* do not provide source code and their work involves several complex yet peripheral algorithms aimed at improving the basic detection accuracy. Our comparison to Ferrari *et al.* (2009) is aimed at demonstrating the usage of clustering for building the colour appearance model, hence we focus on their core ‘best-generic’ approach to colour modelling and eliminate the other factors (as discussed in section 4.4.1). We dub this approach Ferrari-Core to distinguish it from Ferrari’s full proposal.

Three sequences are used to evaluate and compare the performance of the systems, as shown in Figure 4.9. One is the ‘Baseball’ sequence used by Ramanan *et al.* (2007) and two are from the HumanEva data set (Sigal and Black, 2006). The three sequences capture three different people performing different activities. The Baseball video is a sequence of 200 frames that records a pitcher throwing out a ball. The two sequences from the HumanEva data sets are 157 frames (from frame 353 to frame 509) of ‘HE1_S2_Walking_1_C1’ and 251 frames (from frame 220 to frame 470) of ‘HE1_S4_Throw_Catch_2_C1’ respective-

ly, where the first sequence shows an ordinary walking motion and the second sequence records a man throwing and catching a ball where a lateral walking pose does not exist. The approach of [Ramanan *et al.* \(2007\)](#) cannot be applied for such a sequence without re-training a new stylized-pose detector, whereas our approach can be used without the need of re-training. The ground truths in these sequences are labelled manually for evaluating the performance of the tracking systems.



Figure 4.9: Three sequences are used to compare our system with Ramanan’s and Ferreri-Core in this chapter. Walking is short for HE1_S2.Walking_1_C1. Throwing is short for HE1_S4.Throw_Catch_2_C1.

4.4.1 Evaluation and Metrics

For the comparison to [Ramanan *et al.* \(2007\)](#), the author’s code (as provided at his personal website¹) was downloaded and executed. The Baseball sequence used by [Ramanan *et al.* \(2007\)](#) (downloaded from the same source) and the HE1_S2.Walking_1_C1 sequence both satisfy the requirement of Ramanan’s system where a lateral walking pose must exist within the video. In [Ramanan *et al.* \(2007\)](#), temporal smoothing is implemented within their system when evaluating their tracking performance, but the code provided by Ramanan on his personal website does not include the subsystem performing tempo-

¹<http://www.ics.uci.edu/~dramanan/papers/pose/index.html>

ral smoothing. Therefore we implemented temporal smoothing according to the method described in [Ramanan *et al.* \(2007\)](#).

With respect to the approach of [Ferrari *et al.* \(2009\)](#) whose source code is not available, we focus on the core difference between our approach and theirs. Specifically, we re-implement the concept of [Ferrari *et al.* \(2009\)](#) to build a colour appearance model based on the top 10% of the preliminary estimations. This is in contrast to our proposed approach where clustering is used to identify the correct colour appearance. To avoid adding further confounding factors in the comparison, we use this colour model similarly to our clustering-derived model to find the final estimation whose overlaid colours are the closest match to the colour model in the alternative generic detections. Since the colour model is a histogram, we use the Bhattacharyya distance to calculate the error between the model and the generic detections.

Metric-1 As with the evaluation method of [Ramanan *et al.* \(2007\)](#), an estimation of configuration l_m is considered a correct estimation when the majority of pixels within the rectangle l_m are correctly labelled. To formally define the evaluation metric, let $\rho(l)$ be the set of pixels located within the rectangle l and $|\rho(l)|$ be the number of pixels in the set $\rho(l)$. Assuming \widehat{l}_m is the estimated configuration and l_m is the ground truth for part m , \widehat{l}_m is defined as a correct estimation when the following condition is satisfied:

$$|\rho(l_m) \cap \rho(\widehat{l}_m)| \geq \frac{1}{2} |\rho(l_m)|. \quad (4.8)$$

In addition, if a limb is occluded but the tracking algorithm still detects the limb it is considered an erroneous track. Whilst this 50% overlap metric is a fairly relaxed standard, we use it nevertheless, to be consistent with the published results in the existing work.

Metric-2 Another evaluation metrics called Metric-2, suggested by [Sigal and Black \(2006\)](#), are also used. Under this evaluation metrics, human pose (\mathbf{x}) is represented as 20 vectors of two-dimensional body joint positions (20 joints or markers each having X and Y coordinates). The error $D(\mathbf{x}, \widehat{\mathbf{x}})$ between the estimation $\widehat{\mathbf{x}}$ and the ground truth \mathbf{x} is measured as following:

$$D(\mathbf{x}, \widehat{\mathbf{x}}) = \frac{1}{\sum_{i=1}^M K_i(\mathbf{x})} \sum_{i=1}^M (K_i(\mathbf{x}) \times \|m_i(\widehat{\mathbf{x}}) - m_i(\mathbf{x})\|) \quad (4.9)$$

where $m_i(\widehat{\mathbf{x}}) \in \mathbb{R}^2$ is a function that extracts the two dimensional coordinates of the i th joint position, M is the number of the joint positions for each pose and $\|\cdot\|$ is the Euclidean distance. Moreover, $K_i(\mathbf{x}) \in \{0, 1\}$ is a function that determine whether the

i th joint point is visible (not occluded), i.e., if the i th joint point is visible, $K_i(\mathbf{x}) = 1$ otherwise $K_i(\mathbf{x}) = 0$. In addition, Metric-2 can also be used to evaluate a specific body part by assuming that only the joints corresponded to this body part are visible.

According to different accuracy requirement in the different applications, different metrics are chosen to evaluate a system of human pose estimation. The metric-1 is used to roughly evaluate the accuracy of human pose estimation, which is used to generally observe a system. In contrast, the metric-2 evaluates a system in a pixel level, which is appropriate for the application requiring higher accuracy. From research perspective, no particular application requirement is aimed at, so two metrics are used to comprehensively evaluate our system.

4.4.2 Experimental Results – Walking

Experiments are conducted to test the three tracking systems on sequences containing a lateral walking pose: the Baseball sequence and the HE1_S2_Walking.1_C1 sequence. To obtain a comprehensive evaluation for the systems, we first evaluate the system performance using Metric-1 by considering the matches of all body parts including left lower leg (LLL), left upper leg (LUL), right upper leg (RUL), right lower leg (RLL), left lower arm (LLA), left upper arm (LUA), right upper arm (RUA), right lower arm (RLA), head and torso. Next Metric-2 is used to evaluate the three systems by computing the distance of the joints between the estimations and the ground-truth.

4.4.2.1 Comparing with Ramanan’s System based on Metric-1

The screenshot of tracking results and evaluation results based on Metric-1 are shown in Figure 4.10 and 4.11. It is worth noting that the performance of Ramanan’s system shown in Figure 4.11 differs from that reported in Ramanan *et al.* (2007). This is due to the fact that our evaluation metric considers *all* limbs, whereas Ramanan *et al.* (2007) considered only matches of the torso, the best lower leg and the best lower arm, thus ignoring the head and second arm/leg. Ramanan *et al.* (2007) did not consider false detections or missed detections of the second limb. Moreover, Ramanan *et al.* (2007) evaluate their results based on 100 selected frames whereas we evaluate all 200 frames. They explain that their evaluation is done in the specific manner in order to reduce confusion due to occlusions, since only one leg and one arm are always visible, and that there was no need to verify the upper arm/leg since “if ... lower limb estimates are accurate, the upper limbs

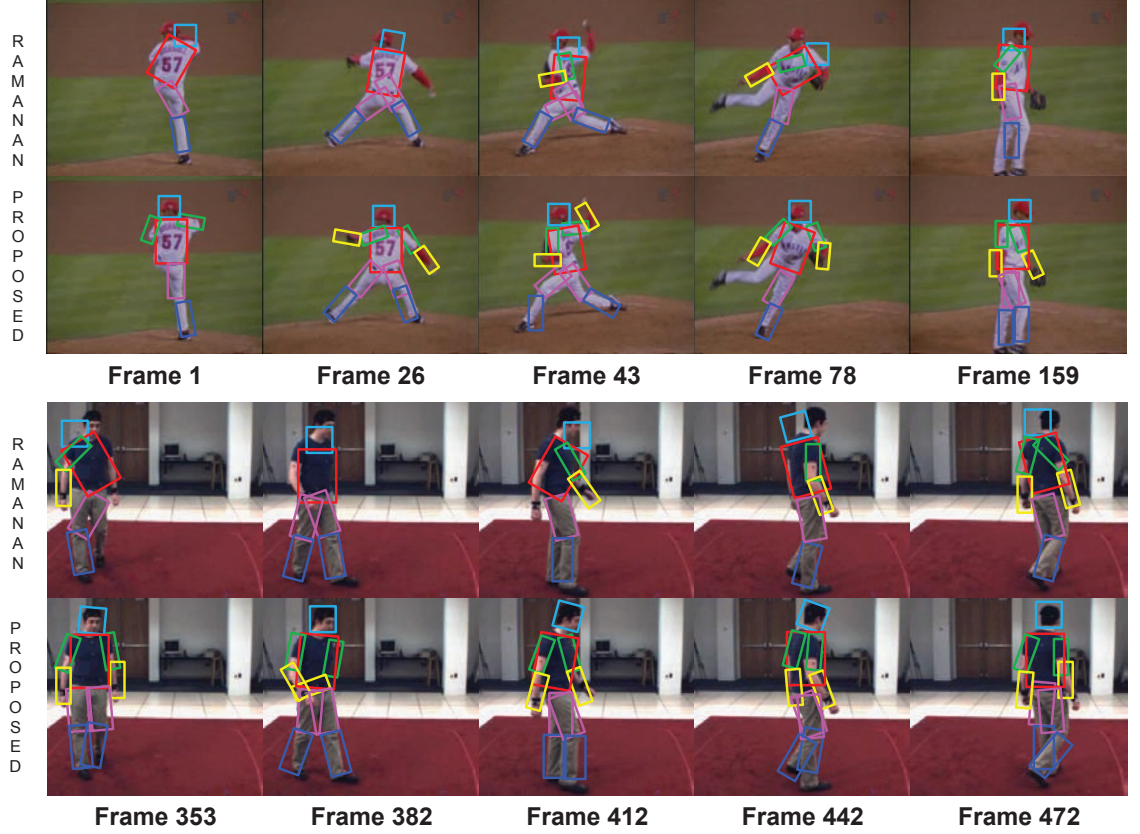


Figure 4.10: The screenshot of tracking results in Baseball and Walking sequences from [Ramanan *et al.* \(2007\)](#)’s system and proposed system

are more or less constrained to be correct” ([Ramanan *et al.*, 2007](#)). We argue that such an evaluation is incomplete – the ability to correctly detect occlusions is an important task, and our results show that the lower arm/leg is not always detected correctly with the upper arm/leg.

As shown in Figure 4.11, in every case the proposed approach outperforms that of Ramanan’s, often significantly. As expected, the tracking performance for the lower arms in the proposed system is much better than that in Ramanan’s system. In detail, for the Baseball sequence, improvement on LLA is from 72.9% to 85.5% and improvement on RLA is from 78.9% to 84%. For the Walking sequence, improvement is from 68.3% to 86.5% on LLA and from 80.5 to 86.5% on RLA. For example, in Figure 4.10, it can be seen that the lower arms in Frame 26, 43 and 78 of the Baseball sequence are not detected by Ramanan’ system but they are detected by the proposed system. This is due to the fact that in the proposed system the specific appearance model is more reliable learned than in Ramanan’s system. In Ramanan’s system, the training samples for learning the

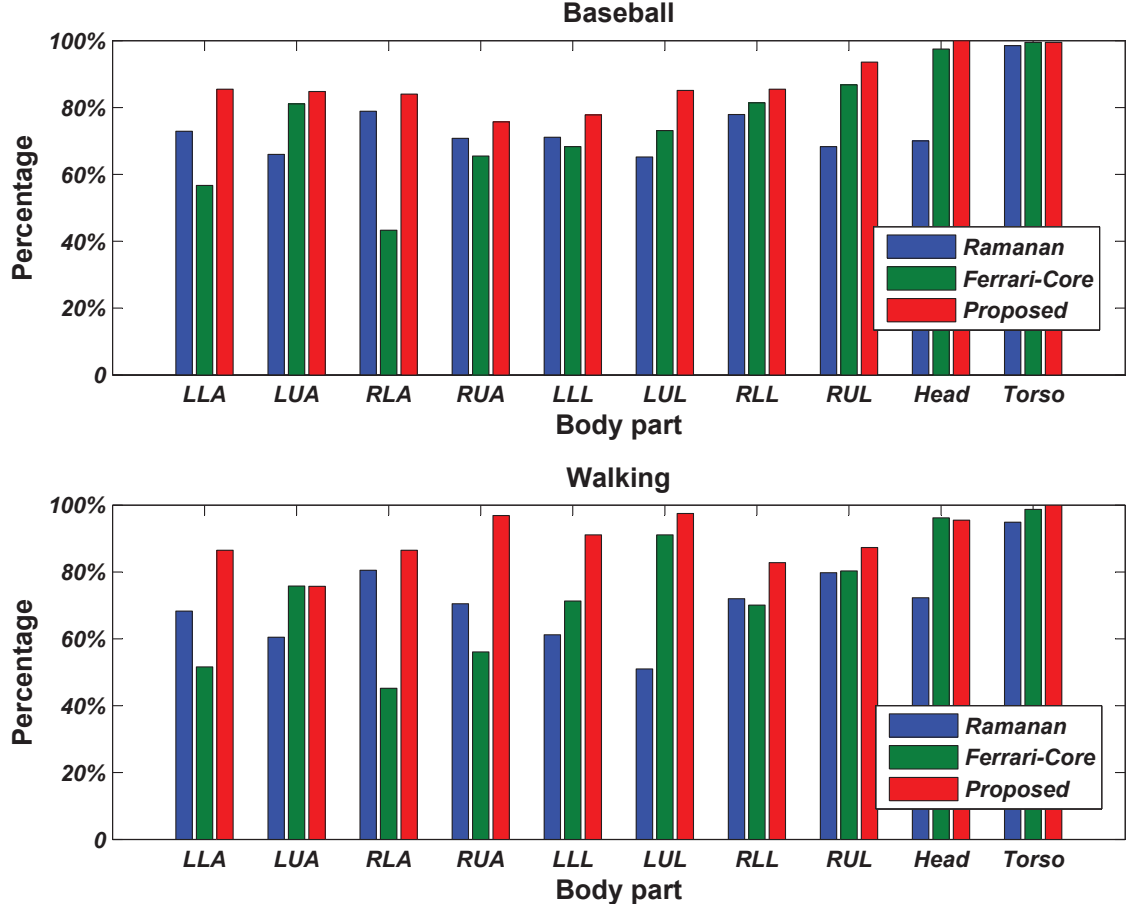


Figure 4.11: Evaluating three systems' tracking performance on Baseball and Walking sequences based on Metric-1.

specific appearance model are extracted from a specific frame (called the training-sample frame) of the whole sequence. For large body parts such as torso and legs, plenty of feature pixels may be obtained and thus a reliable specific appearance model can be learned. However, for smaller body parts such as head and arms, the quantity of the training pixels is possibly insufficient to learn a specific appearance model which is robust to external interference. When the colour appearance for a target is slightly altered due to the change of illumination etc. in other frames, the learned appearance models of small body parts have a more chance to fail than that of large body parts, especially for the frames that are not adjacent to the training-sample frame. In contrast to Ramanan's system, training samples in the proposed system are extracted from multiple frames across all sequence. Thus, the learned appearance model can more effectively handle pose variations and external interferences. It can be noted that in Ramanan's system the accuracy in tracking RLA exceeds the accuracy in tracking LLA, which is close to the proposed system. This

is due to the fact that RLA is often occluded by the torso in the Baseball and Walking sequences. When RLA is invisible, the fact that its specific appearance model is not robust is covered up.

The advantage of the proposed system against Ramanan’s system is also evident in tracking the upper arms as shown in Figure 4.11. In each of the Baseball and Walking sequences, the clothes worn by the person observed has the same colour around the upper arms and the torso, which is very common in the daily clothing of people. If the lower arm is occluded or missed by detection in Ramanan’s system, as shown in Frame 1, 26, 43 and 78 of the Baseball sequence and Frame 353, 382 and 412 of the Walking sequence (Figure 4.10), it is impossible to obtain a correct estimation of the upper arm since a specific (colour) appearance model only is not sufficient to distinguish the appearance of the upper arm from that of the torso. Moreover, even when the lower arm is detected, if the estimation of the torso orientation is incorrect it is also practically impossible to obtain the correct estimation of the upper arm, as shown in Frame 78 of the Baseball sequence and Frame 353 of the Walking sequence (Figure 4.10). With the proposed appearance model combining both the shape and colour features, this issue can be in some extent resolved due to the fact that the proposed system can select a most likely estimation based on shape features when it is confused by colour features.

As shown in Figure 4.10, in comparison with Ramanan’s system, the performance in tracking legs is also improved by the proposed system. To be specific, with respect to the Baseball sequence, the performance for LLL, LUL, RLL and RUL is improved respectively by 6.7%, 19.9%, 7.6% and 25.3%. As for the Walking sequence, the performance for LLL, LUL, RLL and RUL is improved by 29.9%, 46.5%, 10.8% and 7.5%. This improvement benefits from the proposed system appropriately using the combination of generic and specific appearance models. As discussed in Section 2.5.2, in the case that two legs are not well separated, which widely exists in the sequence recording the lateral walking, Ramanan’s system is prone to result in wrong estimation (Please refer to the discussion in Section 2.5.2 for details). Their estimation for a leg usually appears in the middle of two legs (eg: the leg estimations from Ramanan system in Frame 159 of the Baseball sequence and Frame 353, 412 and 472 of the Walking sequence as shown in Figure 4.10). In such cases, generally, only one of the two legs can be detected while another leg is determined as being occluded (invisible) by Ramanan’s system. Such a problem does not usually happen in the proposed system. This is due to the fact that the proposed system is more careful about determining a body part as being occluded (invisible). We assume that every body part is visible at the beginning. Generic detections based on shape features are generated and they are checked by the learned specific appearance model. A body part will not be determined as being occluded (invisible) as long as any generic detection is consistent with

the learned specific appearance model. Therefore, the proposed system is able to detect two legs even though they are not well separated.

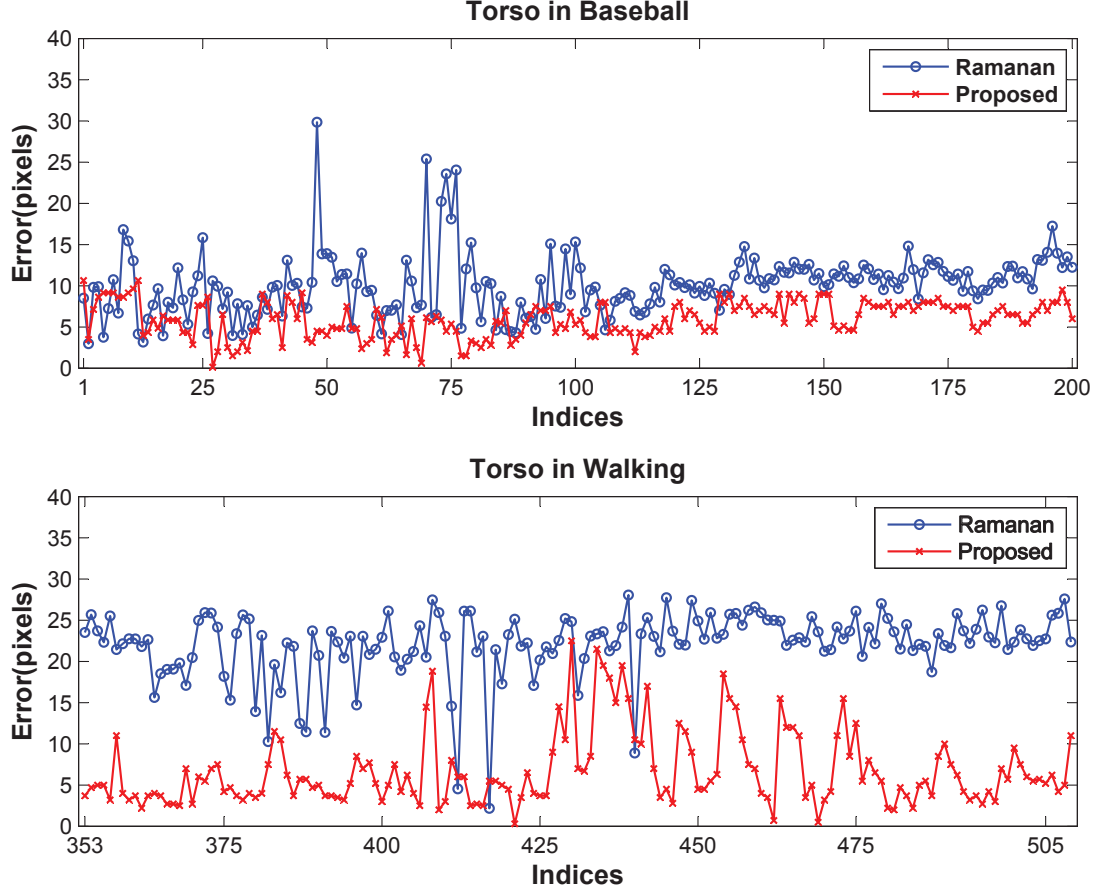


Figure 4.12: Performance in tracking Torso for Baseball and Walking Sequence is evaluated by the Metric-2.

For the torso, the two systems achieve almost the same performance according to Metric-1. It is worth noting that the evaluations based on Metric-1 do not take into consideration the orientation of the body part estimated. Rather, only the coverage of the correct pixels is considered. However, many estimated configurations produced by [Ramanan *et al.* \(2007\)](#) are actually in incorrect orientations despite covering over 50% of the correct pixels (eg: Frame 78 of the Baseball sequence and Frame 353, 412 and 442 of the Walking sequence). Metric-2 is introduced to evaluate the performance in tracking the torso in another way. Unlike Metric-1, incorrect orientation of a body part can be detected by the evaluation based on Metric-2. As shown in [Figure 4.12](#), according to Metric-2, the proposed system is much better than Ramanan’s system in the performance of tracking Torso. Similarly the proposed system outperforms Ramanan’s system in tracking upper arms. An illustration

is given in Figure 4.13 to demonstrate the differences of the two systems in estimating torso configuration. In Ramanan’s approach, the torso pixels are identified from the original image based on the torso colour appearance model, as shown in Figure 4.13 (a). According to the torso pixels identified, multiple candidates have a chance to be chosen as the final torso estimate, as shown in Figure 4.13 (b). Unfortunately, it is impossible to rank these candidates by considering only their colour features. In other words, the decision about which candidate is better cannot be made by comparing only their colour feature response. Thus, Ramanan’s approach is prone to producing an inaccurate torso estimation, as shown in Figure 4.13 (c). In contrast, in the proposed approach, shape features such as the edge feature as shown in Figure 4.13 (d) are included which can be utilized to evaluate the multiple candidates before deriving the final estimation. In this process, the candidates that do not match the shape features well are reduced to a lower rank. The candidate that is consistent with both the colour and shape features are more likely to become the final estimation than the other candidates. Figure 4.13 (e) shows that the best candidate becomes the final estimation in the proposed system. The same could also occur for other body parts besides the torso.

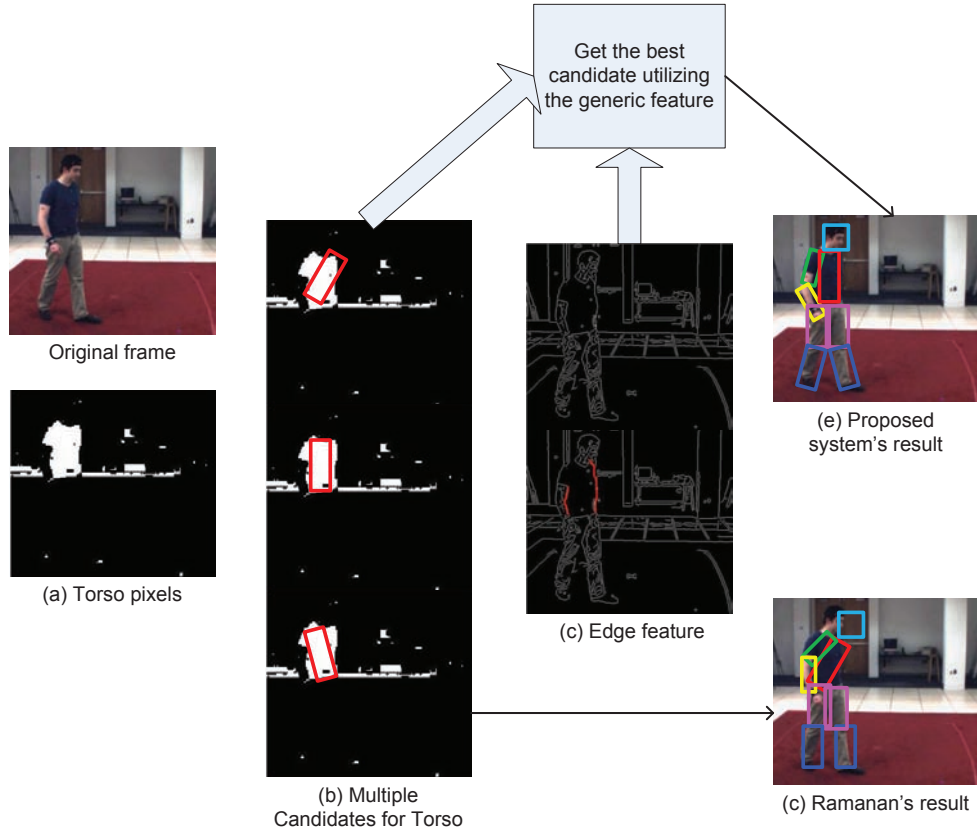


Figure 4.13: Scheme difference between the proposed system and Ramanan’s system.

Finally, as shown in Figure 4.10, the accuracy for head tracking in the Baseball sequence

and the Walking sequence is only around 70% and 72.3% in Ramanan’s approach even though the head is a rather unique and distinct body part. We believe the poor performance is due to the fact that their colour appearance model is built only from a single frame (with the lateral walking pose). Since the area of a head is generally small, it is possible that the number of pixels provided by the training sample is not adequate for the training of an effective appearance model for the head. Even a small number of incorrect pixels could significantly deteriorate the accuracy of the appearance model generated. With the proposed approach learning the colour model is learned from multiple frames, such a problem is very well addressed. Moreover, shape features alone are quite effective for the estimation of head configuration (Andriluka *et al.*, 2009). The proposed system achieves an average head tracking accuracy of over 95%.

4.4.2.2 Compare with Ferrari-Core’s System based on Metric-1

When comparing our approach to that of the ‘best-generic’ method of Ferrari-Core, the main difference is that the proposed approach uses clustering to select the largest subset of preliminary generic detections that are consistent with one another. Ferrari’s assumption of the most-likely preliminary generic detections representing the accurate matches for the limb detection turns out to be less than ideal. In addition, unlike Ferrari-Core that merges the generic and specific appearance models via averaging, the proposed approach makes the final estimations conform to both the generic and specific appearance models. The proposed clustering-based approach outperforms Ferrari-Core approach, usually significantly, in all but two instances: the HumanEva sequence for the left upper arm and head. In these two cases, the difference in performance is marginal (0.1% and 0.7% respectively), indicating that for these parts the most-likely preliminary estimates happened to be accurate and so the colour model built would be similar to that of our clustering approach. However, for other, more difficult, parts such as the lower arms and legs, Ferrari-Core’s reliance on the most-likely estimates being accurate has been proven to be flawed. In such cases, our clustering approach comprehensively outperforms Ferrari-Core approach by wide margins.

4.4.2.3 Evaluation based on Metric-2

The evaluation results based on Metric-2 for three systems in the Baseball sequence and the Walking sequence are shown in Figure 4.14. It can be clearly seen that the accuracy of the proposed system is more than that of Ramanan’s system or Ferrari-Core system,

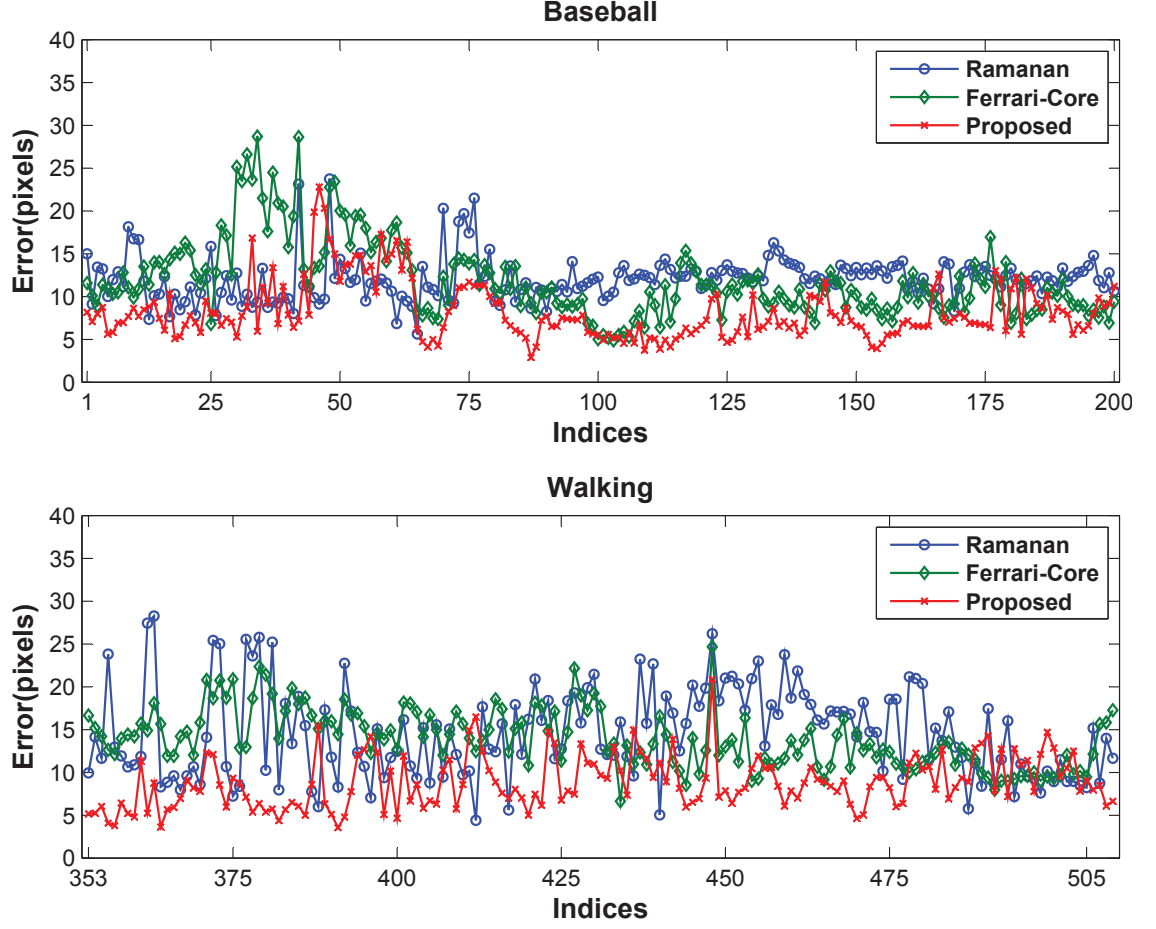


Figure 4.14: The evaluation results based on Metric-2 for three systems in Baseball sequence and Walking sequence. Error is average error across all body parts.

which further demonstrates that the proposed system outperforms Ramanan’s system and Ferrari-Core system. More specifically, for the Baseball sequence, the proposed system achieves more accurate estimations than Ramanan’s system and Ferrari-Core system in 83% and 87% of all frames respectively. The average error for the proposed system in Baseball sequence is 8.2 pixels distance, whereas they are 12.0 and 11.8 pixels distance from Ramanan’s system and Ferrari-Core system respectively. For the Walking sequence, the proposed system outperforms Ramanan’s and Ferrari’s systems in 83.4% and 84.7% of all frames respectively. The average error is 8.8 pixels distance in the proposed system, while they are 14.5 and 13.9 pixels distance in Ramanan’s and Ferrari-Core systems respectively. As analyzed in Section 4.4.2.1 and 4.4.2.2, this improvement benefits from reliably learning a specific appearance model and combining both the generic and specific appearance models in an appropriate way.

4.4.3 Experimental Results – Non-walking Sequences

This section examines the case of video sequences containing motions that do not contain a lateral walking pose. Such sequences are very common, but Ramanan’s approach fail to work for them. When the required pose does not exist, Ramanan’s system requires identifying another ‘known-pose’ that exists in the video sequence and re-train the system, in order to initialize tracking. In comparison, our approach does not rely on any specific poses or motion models – as long as the person to be tracked is detailed enough for low-level edge-based shape detectors to work, tracking can be accomplished. The screenshots of tracking results on the Throwing sequence by the proposed system is shown in Figure 4.15. The experimental results based on Metric-1 and Metric-2 for tracking the Throwing sequence that contains no lateral walking pose are shown in Figure 4.16 and Figure 4.17.

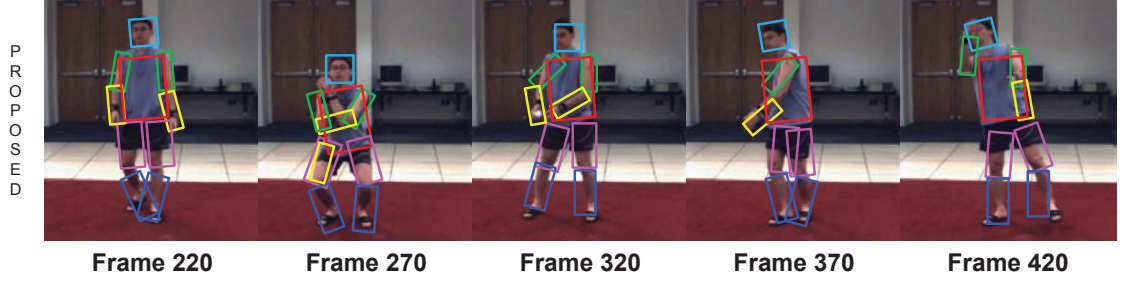


Figure 4.15: Screenshot of tracking results on Throwing sequences

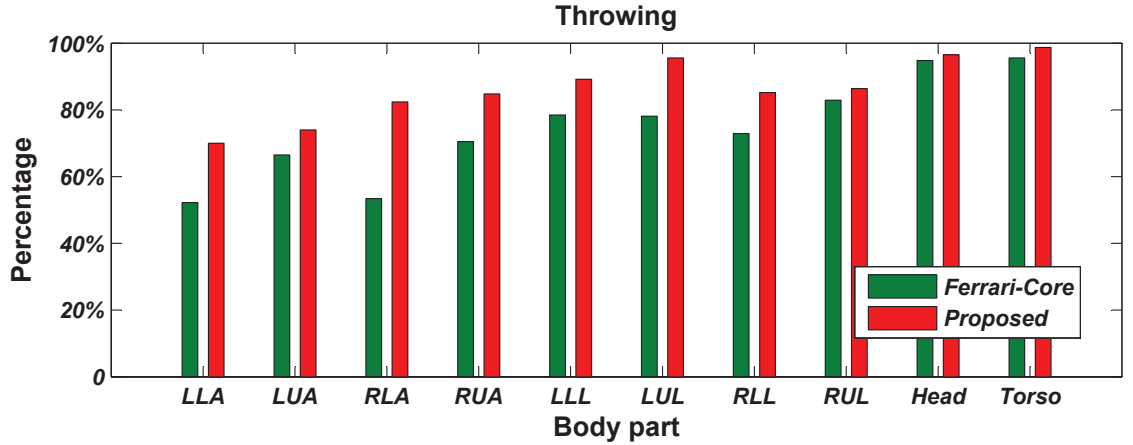


Figure 4.16: The evaluation results based on Metric-1 for two systems (Ferrari-Core and proposed) in Throwing sequence that contains no lateral walking pose.

Unlike Ramanan’s method which uses an appearance model based only on colour features,

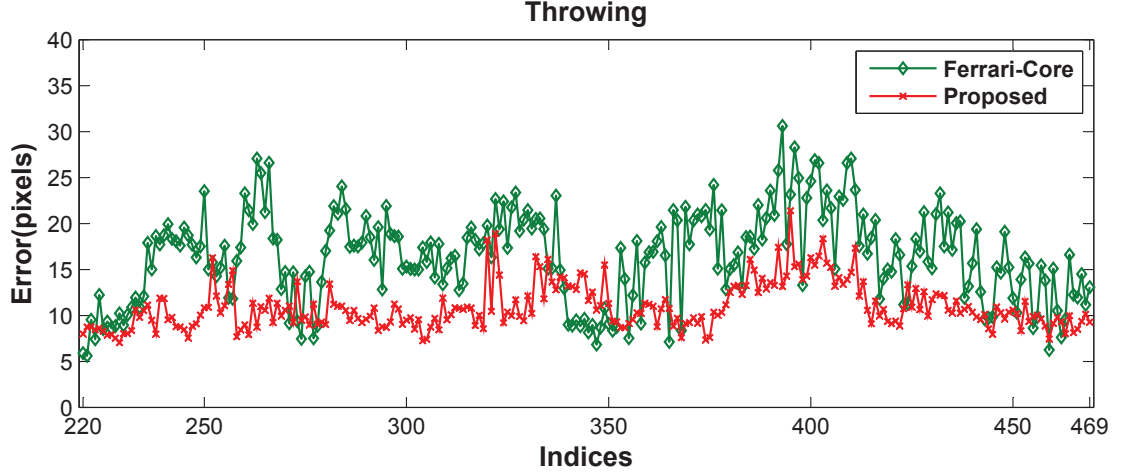


Figure 4.17: The evaluation results based on Metric-2 on two systems (Ferrari-Core and proposed) in Throwing sequence.

both Ferrari-Core system and the proposed system make use of an appearance model combining both the shape and colour feature. When comparing the proposed system with Ferrari-Core system, the main difference is that clustering is utilized in the proposed approach to select the largest subset of preliminary generic detections that are consistent with one another in colour appearance. Based on the largest subset, the specific appearance model is built in the proposed system. Unlike the proposed system, Ferrari-Core system directly uses the best generic detections to learn a specific appearance model based on the assumption of the most-likely preliminary generic detections representing accurate matches for the limb detection, which turns out to be less than ideal. As expected, the evaluation results based on Metric-1 and Metric-2 further demonstrate that the proposed system outperforms Ferrari-Core system.

4.5 Chapter Summary

In order to estimate 2D human configurations over time based on monocular view, two recent bottom-up approaches based on learning a specific appearance model have achieved good performances, but some issues exist in the two approaches. In [Ramanan *et al.* \(2007\)](#), the specific appearance model cannot be learned when the lateral walking pose does not exist in a video sequence, and the reliance on using a single frame to learn the colour appearance means the specific appearance model may not be robust with regard to lighting variations throughout the video sequence. In [Ferrari *et al.* \(2009\)](#), the effectiveness of the

specific appearance model depends on the assumption that generic detection likelihood is representative of accuracy between frames and so that the top few optimal estimations will be the most accurate. However, the assumption has been experimentally proved to be unreliable in Chapter 3. In addition, in order to improve the accuracy of tracking by using both generic and specific appearance models, Ferrari *et al.* (2009) chooses to merge the generic and specific maps via averaging, which will not enforce that both models agree on the final estimations.

Aimed to overcome these issues in Ramanan *et al.* (2007) and Ferrari *et al.* (2009), this chapter presents a system to estimate unrestricted 2D human configurations over time. We cluster generic pose estimations in terms of their colour to identify the subset of optimal estimations that are accurate, based on the assumption that correct estimations have a similar colour appearance. Then a specific appearance model is learned using these identified correct optimal estimations. Finally, for frames whose optimal estimations do not cluster as ‘correct’, the top N most likely generic estimations for each incorrect frame is filtered to find candidate postures that conform to the correct cluster. This approach obtains the pose estimations that agree with the evidence of both the generic and specific appearance models. Three sequences are tested against proposed system in comparison with Ramanan *et al.* (2007) and Ferrari *et al.* (2009). The experimental results show that proposed system outperform Ramanan *et al.* (2007) and Ferrari *et al.* (2009), often significantly.

Chapter 5

Building an Uncontaminated Specific Appearance Model

To overcome the shortcomings of the approaches proposed by [Ramanan *et al.* \(2007\)](#) and [Ferrari *et al.* \(2009\)](#), the approach proposed in Chapter 4 avoids naïvely choosing the maximum posterior score postures as correct estimations. Instead, rough poses are estimated using an edge-based generic limb detector and these estimations are then clustered based on their *colour*, with the largest cluster (per-limb) indicating the human’s specific appearance model. The proposed approach of using both generic edge and specific colour-based features has been shown to achieve better performance than both [Ramanan *et al.* \(2007\)](#) and the core approach of [Ferrari *et al.* \(2009\)](#).

There is some room for improvement in the proposed approach. A most significant problem is as follows, which is a problem for all approaches that need to learn a specific appearance model using the generic estimations from a generic pose detector. In order to learn the specific appearance model, training samples are selected from the generic estimations of the generic human pose detector. Even if the training samples are accurate estimations in the sense of the ground-truth, non-target pixels are unavoidably included in the training samples thus contaminating the learned specific appearance model. This is due to the fact that the generic estimations of a body part (even if they are correct estimations) would not cover its area in the image exactly, especially since the precise shape, size and boundaries of the tracked person’s body parts are unknown.

In this chapter, in order to overcome such a problem, we propose a method to significantly reduce non-target pixels from training samples in order to build specific appearance models that are not (or less) contaminated by non-target (ie: non-body-part) noise. The general approach discussed in Chapter 3 is employed where a generic pose detector based on the pictorial structure is first used to roughly estimate human poses in each frame of a video sequence. A set of the preliminary estimations for a body part across all frames (Section 3.4.1) is obtained. Following the approach described in Section 4.2, clustering is used to divide the preliminary estimations into several clusters according to their colour histogram features, with the largest cluster indicating correct estimations among the set of

preliminary estimations. Moreover, it is known that these preliminary estimations will be contaminated with pixel colours from the background. Thus, at this point we diverge from the approach described in Chapter 4. Rather than learning a specific appearance model from the preliminary estimations in the largest cluster directly, we further analyze the colours in this cluster in order to remove non-target pixels. Only the identified target pixels in the correct estimations are extracted to learn an uncontaminated specific appearance model. Tracking is implemented based on this learned uncontaminated specific appearance model.

The rest of the chapter is organized as follows: Section 5.1 provides an overview of the improved tracking system and outlines the process of learning the uncontaminated colour appearance models. Section 4.2 discusses how to automatically build an uncontaminated colour appearance model based on mean shift and the assumption that the central part of a detection tends to overlap the true limb whereas non-target pixels tend to exist near the boundary. Section 4.3 presents the implementation of tracking based on the learned uncontaminated colour appearance model. Section 4.4 presents the experimental results and discussions.

5.1 System Overview

Figure 5.1 (a) shows an overview of our improved 2D human pose tracking system. The system consists of four major components:

1. Generic human pose detector
2. Colour histogram clustering
3. Building uncontaminated appearance model
4. Detecting human pose based on uncontaminated appearance model

The first two components have been discussed in detail in Chapter 3 and Chapter 4. The last two components are described in details in this chapter.

Using the generic human pose detector and colour histogram clustering described in Chapter 3 and 4, the correct (though contaminated) estimations from the preliminary estimations for a body part across all frames are identified. The pixels specified by these correct but contaminated estimations usually include both the target pixels and the non-target

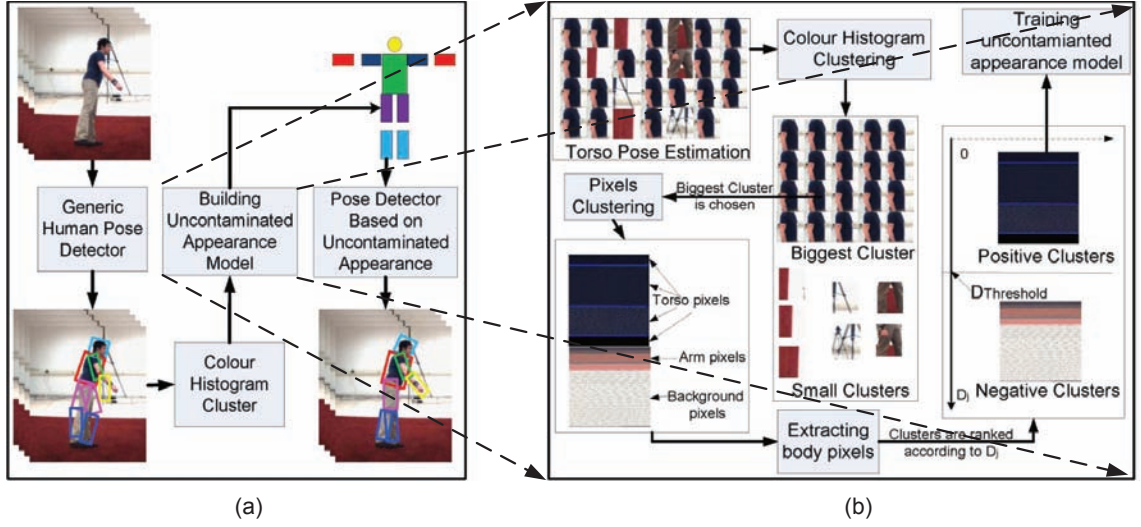


Figure 5.1: (a) Overview of the improved human pose tracking system (b) Visualization of how we build an uncontaminated specific appearance model for human pose tracking based on the results from generic human pose detection and the results of colour histogram clustering. The process of training the torso appearance is used to illustrate our approach.

pixels. In order to build an uncontaminated specific appearance model, we need to extract the target pixels by removing the non-target pixels. The process of extracting target pixels consists of two parts: pixel colour clustering and target pixels extraction, as shown in Figure 5.1 (b). As described in Chapter 4, mean shift (Comaniciu and Meer, 2002) is chosen to look for the pixel clusters. After pixel clustering, pixels with similar colour are gathered to generate several clusters. To separate the positive clusters containing target pixels from the negative clusters containing non-target pixels, two assumptions are made:

- In the correct estimations, the number of target pixels exceeds the number of non-target pixels. This assumption is reasonable given that the correct estimations are typically representative of the correct positions of the body parts, overlapping the ground truth by at least 50%.
- If a bounding box is used to bound all pixels in a correct estimation, most of the target pixels would be located along the central axis of the bounding box and most of the non-target pixels would fall along the border of the bounding box. Thus the clusters whose pixels come from the central area are considered part of the body and used to build an accurate appearance model, and other clusters are discarded. Again the assumption is reasonable due to the fact that the correct estimations, although contaminated, are good representation of the body parts in the images.

Based on these two assumptions, the positive clusters can be separated from the negative clusters. The pixels in the positive clusters are then used to learn an accurate/uncontaminated appearance model. Using the pictorial structure based on the learned accurate appearance model, tracking can be achieved more accurately by detecting human pose in each frame.

5.2 Recap of Prerequisite Approaches

Our proposed approach in this chapter is built based on the human pose detector described in Chapter 3 and the approach of colour histogram clustering proposed in Chapter 4. The relevant aspects of the two algorithms are recapped as follows. In Chapter 3, a generic pose detector is built based on the pictorial structure. When estimating the human configuration in frame t , the maximum posterior estimation s_m^t for body part m is chosen as the optimal estimation for that body part in frame t . Since the optimal estimation will be further processed to derive a more accurate estimation, it is also called a preliminary estimation. All preliminary estimations for body part m across all frames form a set $S_m = \{s_m^t\}_{t=1}^T$, where T is the number of frames in a sequence. The specific appearance feature of a preliminary estimation is represented by its corresponding colour histogram. Assuming that the majority of the preliminary estimations in set S_m are correct in the sense of ground truth and correct estimations have similar colour features (i.e.: consistent over time), preliminary estimations in the set S_m are clustered based on their colour histogram features, with the largest cluster indicating estimations with correct appearance. The preliminary estimations in set S_m existing in this largest cluster form a set $U_m = \{s_m^{t_n}\}_{n=1}^K$, where $1 \leq K \leq T$, $1 \leq t_n \leq T$ and $U_m \subseteq S_m$. Discussions in the following two sections refer to estimations in set U_m .

5.3 Removing the Non-target Pixels from Contaminated Correct Estimations

As discussed in Chapter 4, in order to learn a specific appearance model for body part m , the appearance features of the preliminary estimations in set U_m , i.e., the preliminary estimations in set S_m that exist in the largest cluster, are extracted to represent the appearance features of the correct estimations. However, since the precise shape, size and boundaries of the body part are unknown, a preliminary estimation, even if it is considered correct, cannot perfectly cover the area of a body part in the image, i.e., both the target pixels and non-target pixels co-exist in the area specified by a preliminary estimation. The

non-target pixels may either be from the background of the scene or from other body parts of the tracked person. Therefore, the preliminary estimations in set U_m can also be called *contaminated correct estimations*. As shown in Figure 5.2 (a), the left figure is an example of a contaminated correct estimation from the Baseball sequence. To provide a better view, a part of the figure is enlarged in the right figure. It can be clearly seen here that both the target pixels (body part pixels) and non-target pixels (background pixels) are included in the correct estimations. Specifically, the pixels enclosed by black dotted line belong to the body pixels whereas the pixels outside the dotted line are the background pixels. Note that in this example the non-target pixels are from the background. Sometimes they can also come from other body parts, as shown in Figure 5.2 (b). Here the preliminary estimation for the right lower arm is shown. The pixels inside the white dotted line belong to the right lower arm (target body part) and the pixels outside the white dotted line are from either the torso or the right upper leg (non-target body parts). Both the background pixels and pixels from non-target body parts are called the non-target pixels. Only the pixels from the targeted body part are called the target pixels.

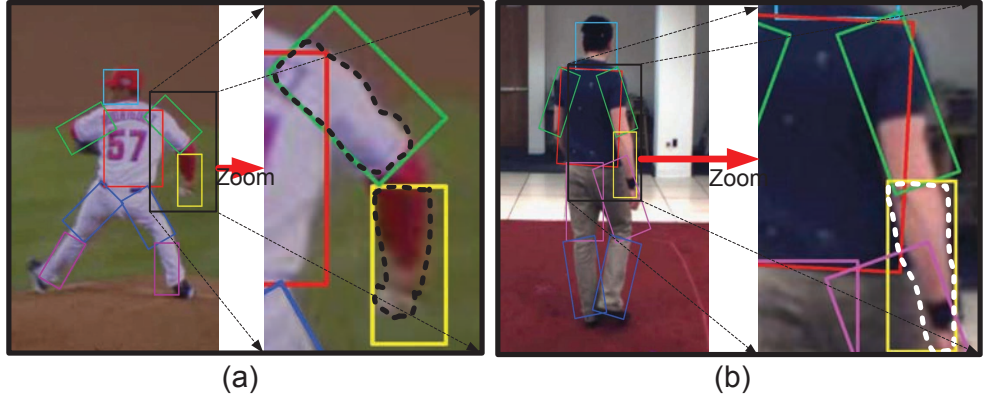


Figure 5.2: Two examples show that both target pixels and non-target pixels co-exist in correct estimations. Figure (a) shows that non-target pixels may come from the background. Figure (b) shows that non-target pixels may also come from the non-target body parts.

If a specific appearance model is built using the pixels (including both the target and non-target pixels) specified by the contaminated correct estimations, such as the colour appearance model described in Chapter 4, its effectiveness will likely be degraded due to the inclusion of non-target pixels.

In this chapter, our aim is to learn an accurate/uncontaminated specific appearance model, which is not (or less) contaminated by the non-target pixels, even when it is still based on the contaminated correct estimations. To achieve this goal, the first problem we need to

solve is how to separate the non-target pixels from the target pixels in the contaminated correct estimations.

5.3.1 Basis of Removing Non-target Pixels

Due to the fact that the contaminated correct estimations will not perfectly cover the area of a body part in the image, if all the pixels specified by the contaminated correct estimations are directly used for learning a specific (colour) appearance model, the non-target pixels will be inevitably involved in the training samples thus contaminating the learned appearance model. In order to learn an accurate specific (colour) appearance model using the contaminated correct estimations, we need to develop an algorithm for identifying the target pixels in the contaminated correct estimations. These identified target pixels can be used to learn an accurate specific appearance model. It is therefore necessary to explore the characteristics that are able to effectively distinguish the target pixels from non-target pixels.

5.3.1.1 Target Pixels Outnumber Non-target Pixels

As shown in Figure 5.2, although the target pixels and non-target pixels co-exist in the area specified by the contaminated correct estimations, the majority of these pixels belong to the target body part. This suggests that the number of target pixels exceeds the number of non-target pixels in the contaminated correct estimations. In order to explore whether this characteristic is consistently shared across the contaminated correct estimations, an analysis is conducted in the Walking sequence from the HumanEva dataset (Sigal and Black, 2006), as shown in Figure 5.3. Sequences from this dataset are representative of the type of videos that we seek to process. Moreover, the HumanEva dataset has two features: 1) the images in this dataset are of good quality, and 2) there is no severe motion-blur appearing in the images, which produces a different kind of colour contamination that we do not seek to address in this thesis.

After the ground truth for this dataset is manually marked, the target pixels and non-target pixels can be automatically identified in each estimation. Given all the contaminated correct estimations for a body part, the number of the target pixels and non-target pixels in each contaminated correct estimation can be computed. The total number of the target pixels and the total number of the non-target pixels in all contaminated correct estimations for a body part can also be computed. We can conduct an analysis to compare the total

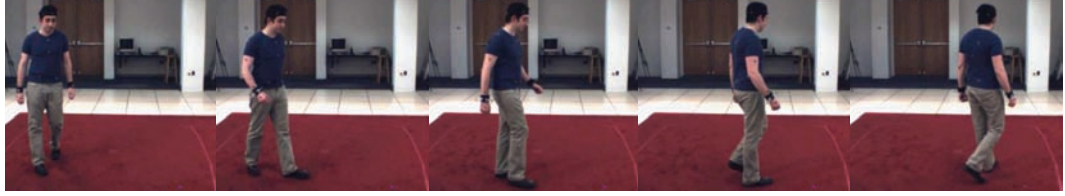


Figure 5.3: Walking sequence (HumanEva.I_Walking_S2) from HumanEva dataset (Sigal and Black, 2006).

number of the target pixels with that of the non-target pixels in all contaminated correct estimations for each body part, i.e., left lower leg (LLL), left upper leg (LUL), right upper leg (RUL), right lower leg (RLL), left lower arm (LLA), left upper arm (LUA), right upper arm (RUA), right lower arm (RLA), head and torso.

The result of analysis for the Walking Sequence is shown in Figure 5.4 where the target pixels and non-target pixels are represented by the dark green label and the light yellow label respectively. It can be clearly seen that the total number of target pixels exceeds the number of the non-target pixels for every body part. Specifically, for head, the percentage of the target pixels in all contaminated correct estimations is about 80% and correspondingly the percentage of non-target pixels is about 20%. For LUL, RUL, Torso, the target pixels account for about 70% and the non-target pixels account for about 30%. In comparison with other body parts, LLL contains more non-target pixels and less target pixels, but the percentage of the target pixels is still over 60%. Although individual detections might not always consists of more target pixels than non-target pixels, pixel colour clustering is done across all detections, so it is not a concern.

In ideal conditions where the colour is invariant to the illumination and the body part is single-coloured, it is reasonable to assume that the target pixels for a specific body part in different frames are of same colour. For instance, torso pixels have a specific colour, even if they are located in different frames. The target pixels will therefore be gathered to a single cluster since they have the same colour. If all the target pixels are gathered into one cluster and the non-target pixels are gathered into other clusters, the characteristic that the total number of target pixels exceeds that of the non-target pixels can be utilized to separate the target pixels from the non-target pixels for each body part by selecting the pixels in the largest cluster as the target pixels, similar to the approach describe in Chapter 4 when clustering colour histograms.

However, such a situation is unrealistic. In particular, some body parts are often multi-coloured such as the upper arms in the Walking sequence as shown in Figure 5.3. Given

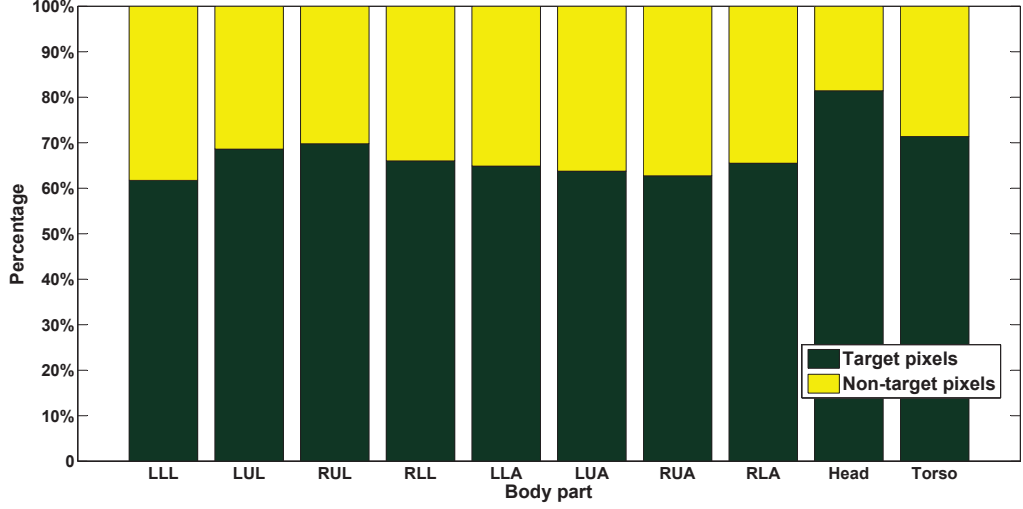


Figure 5.4: The comparison in the quantities of target pixels and non-target pixels specified by contaminated correct estimations for each body part in the Walking sequence.

that a body part can be multi-coloured, even if the target pixels can be clustered by utilizing their colour feature, the target pixels will be divided into multiple clusters, i.e., all target pixels cannot be gathered to a single cluster because they might have different colours. Therefore, the largest cluster may not represent the target pixels even if the number of target pixels exceeds the number of non-target pixels. Thus we cannot use the approach for single-colour body part by selecting the largest cluster to identify target pixels. An alternative algorithm must be developed which is detailed in the following sections.

5.3.1.2 Body Pixels Located In the Central Area

Beside the fact that the target pixels outnumber the non-target pixels in the contaminated correct estimations, another characteristic of the contaminated correct estimations can be derived through observation, i.e., the target pixels are more likely to appear along the central axis. Specifically, when a contaminated correct estimation is specified by a bounding box in an image, as shown in Figure 5.5 (a), the target pixels are more likely to appear in the green area which is closer to the central axis than the white area which is closer to the boundary.

In order to demonstrate this characteristic, the Walking sequence from HumanEva datasets (Sigal and Black, 2006) is again used that includes precise limb ground truth information. Given all contaminated correct estimations for a body part, the number of target pixels and

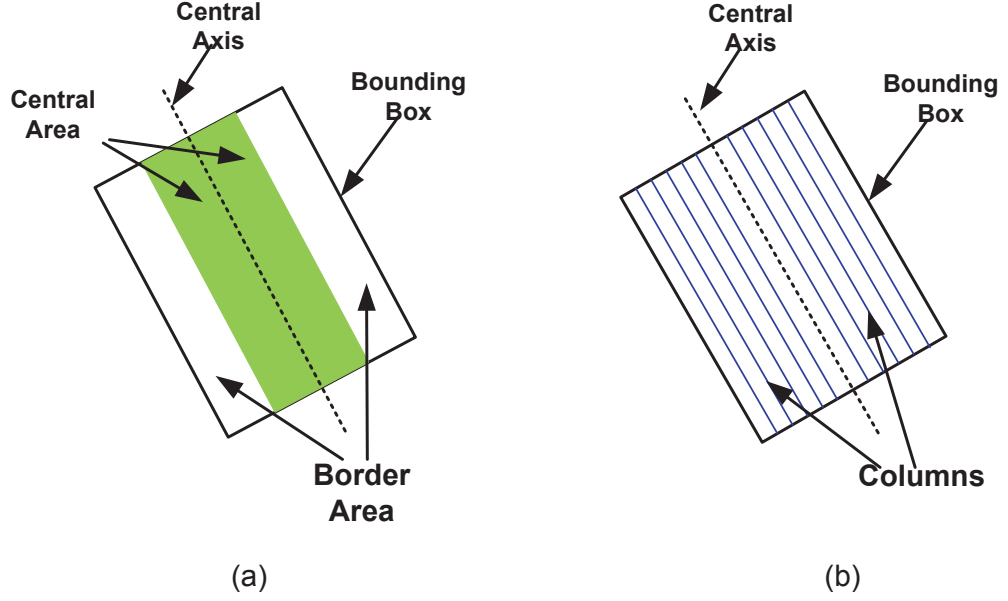


Figure 5.5: (a) An estimation is represented by a bounding box in an image. The green area represents the central area and the white area represents the border area of the bounding box. (b) A figure to demonstrate the columns of a bounding box.

the number of non-target pixels are separately computed in each column of bounding box that is parallel to the central axis as shown in Figure 5.5 (b) in each contaminated correct estimation. The total numbers of the target pixels and the non-target pixels appearing in each column of bounding box across all contaminated correct estimations can then be computed. The total number of target pixels can now be compared with that of non-target pixels in each column. If the total number of target pixels is more than that of non-target pixels in the columns that are close to the central axis, it can be concluded that the target pixels are more likely to appear near the central area than the outer border.

Figure 5.6 shows the percentage of the target pixels and non-target pixels in each column for body parts LLL, LUL, LLA, RUA, Head and Torso. Due to the symmetrical structure of human body, the remaining body parts have similar characteristic to their opposite body parts. Overall, the total number of the target pixels are much more than that of the non-target pixels in the columns close to the central axis for each body part. Specifically, among the columns whose perpendicular distance to central axis is 1, the percentage of target pixels is over 70% while the percentage of non-target pixels is below 30%, even for body part LLL, which contains more non-target pixels than other body parts as shown in Section 5.3.1.1. Please note that the percentage of target pixels cannot reach 100% even in the central axis. This is because the precise shape of human body is unknown and the bounding boxes used to approximate the body parts is usually slightly larger than

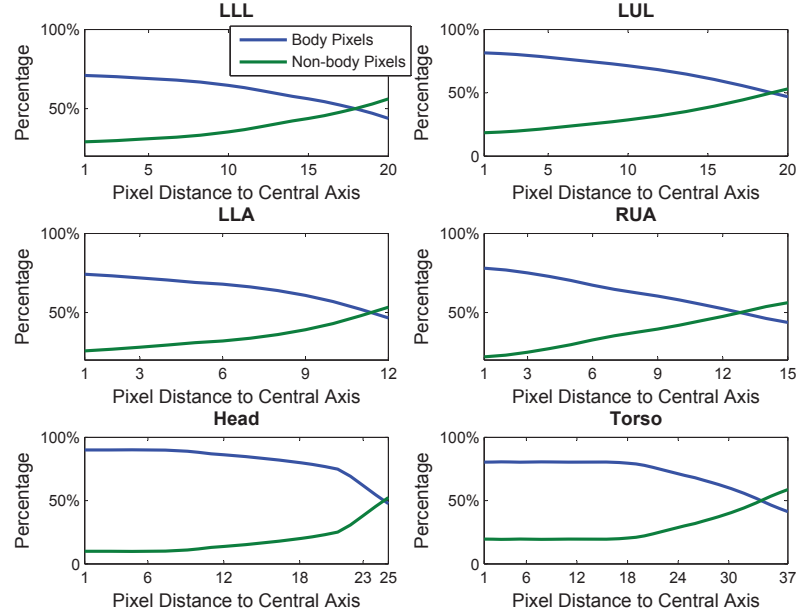


Figure 5.6: The percentage of target pixels and non-target pixels for each column of bounding box across all correct estimations in body parts LLL, LUL, LLA, RUA, Head and Torso.

the real human body parts. As the distance of the column with regard to the central axis increases, although the percentage of the target pixels decreases and the percentage of the non-target pixels increases, the number of the target pixels still obviously exceeds the number of the non-target pixels until the outer border is reached. This suggests that the average perpendicular distance of the target pixels to the central axis is always less than that of the non-target pixels, which could possibly be used as a useful characteristic to separate target pixels from non-target pixels. Table 5.1 shows the average perpendicular distances of the target pixels, all pixels and the non-target pixels to the central axis across all contaminated correct estimations for each body part. It clearly shows that the average distance of the target pixels is always less than the average distance of all pixels for each body part, while the average distance of the non-target pixels is always more than that of all pixels.

This characteristic can be utilized to remove the non-target pixels for a body part after clustering the pixels of all contaminated correct estimations for that body part in a colour space. The average distance with regard to the central axis for all pixels in each cluster can be computed. The clusters whose average distances exceed the average distance of all pixels can be discarded as they contains non-target pixels. Such a characteristic is

Body Part	Average Distance to Central Axis		
	Target pixels	All pixels	Non-target pixels
LLL	9.7528	10.5	11.6929
LUL	9.6282	10.5	12.3746
RUL	9.5215	10.5	12.7192
RLL	9.6305	10.5	12.1678
LLA	6.0706	6.5	7.2786
LUA	7.2473	8	9.2385
RUA	7.1564	8	9.3715
RLA	6.1735	6.5	7.2154
Head	12.1361	13	16.7453
Torso	17.3860	19	22.9980

Table 5.1: Average distance to central axis for target pixels, all pixels and non-target pixels specified by correct estimations.

especially important for multi-colour body parts. As discussed in Section 5.3.1.1, for a multiple-colour body part, the multi-colour pixels will be separated into multiple clusters. Using the method discussed in in Section 5.3.1.1, the largest cluster would be chosen to represent the correct appearance. For a multi-colour body part, the largest cluster only represents one of many colours for the pixels in that body parts hence may not represent the correct appearance. Instead, only a part of the target pixels are identified, even in an ideal case. If the characteristic discussed in this section is used, each cluster will be further analyzed by computing their average distance from the central axis. The clusters whose average distances are less than the average distance of all pixels will be accepted as containing the target pixels. Multiple clusters can be found which could represent multiple colours in a body part.

5.3.2 Extracting Target Pixels via Pixel Analysis

In the previous section, we provide some conceptual and intuitive explanations of our algorithm in building an uncontaminated specific appearance model. In this section, formal notations and definitions will be provided to develop the algorithm in details. Some experiments are incorporated with the description to make the explanation easier to follow. The experiments are conducted on the Combo sequence from the HumanEva dataset (Sigal and Black, 2006), as shown in Figure 5.7.

In order to identify the target pixels for body part m , we need to analyze the pixels specified by the *contaminated correct estimations* in the set U_m . Note that $U_m = \{s_m^{tn}\}_{n=1}^K$ is a

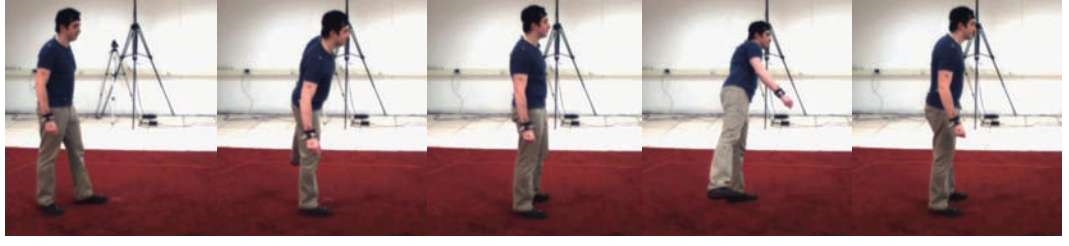


Figure 5.7: Combo sequence (Human_Eva.I.S2.Combo.2.C2) from HumanEva dataset (Sigal and Black, 2006).

subset of set S_m consisting of all preliminary estimations for body part m in a sequence, as discussed in Section 5.2. Since these pixels will be analyzed utilizing their colour features, we need to first transform these pixels into colour feature vectors. Formally, an image patch $a_m^{t_n}$ in frame t_n can be specified by a contaminated correct estimation $s_m^{t_n}$. Each pixel in the image patch $a_m^{t_n}$ corresponds to a colour feature vector in the Lab colour space. All pixels in the image patch $a_m^{t_n}$ can be transformed to a set $Y_m^{t_n}$ of colour feature vectors. Let

$$Y_m^{t_n} = \{K_i\}_{i=1}^{w \times h}, \quad (5.1)$$

where K_i represents a colour feature vector that corresponds to a pixel in the image patch $a_m^{t_n}$, and w, h represents the width and height of the image patch $a_m^{t_n}$. Let

$$Y_m = \bigcup_{n=1}^K \{Y_m^{t_n}\}. \quad (5.2)$$

Thus, all pixels specified by contaminated correct estimations in the set U_m are transformed colour feature vectors in the set Y_m . When Equation (5.1) is substituted into Equation (5.2), the set Y_m can also be denoted as

$$Y_m = \{K_i\}_{i=1}^L, \quad (5.3)$$

where $L = w \times h \times K$ and K is the size of the set U_m .

5.3.2.1 Pixel Clustering

In order to identify the target pixels based on their distances to the central axis (as discussed in Section 5.3.1.2), we need to divide the pixels specified by the contaminated correct estimations in set U_m into clusters according to their colours. After all the pixels specified by the contaminated correct estimations in set U_m are represented by colour feature vectors in set Y_m , pixel clustering can be implemented by clustering these pixels using their colour feature vectors.

Let $Y_m = \{K_i\}_{i=1}^L$ be a set of colour vectors to be modelled. Following the approach of Abd-Almageed and Davis (2007), mean shift (Comaniciu and Meer, 2002) is applied to obtain the modes of the colour vectors. However, unlike Abd-Almageed and Davis (2007), there is actually no need to re-partition the data via using the positive-definite Hessian in the vicinity of the modes as a substitute for the mode's covariance. Instead, since our estimates are reasonably good and are representing single body parts rather than the whole body, the complexity of the appearance is lower and clusters tend to be roughly Gaussian. Hence the partitioning provided by mean shift is used directly rather than introducing the additional re-partitioning step that Abd-Almageed and Davis (2007) performs.

After the process of mean shift clustering, a set of l modes, i.e., $Y_{m_c} = \{K_{c_j}\}_{j=1}^l$ is generated, which represents the local maxima points, where $l \ll L$. Thus each vector in Y_m is attached to a mode in set Y_{m_c} (according to the mean shift) and the set of colour vectors Y_m is partitioned as

$$Y_m = \bigcup_{j=1}^l Y_m^j \quad (5.4)$$

where Y_m^j is one of all clusters which corresponds to the mode K_{c_j} in the set Y_{m_c} of l modes.

Figure 5.8 gives two examples of the pixel clustering from the Combo Sequence, one is for the left lower leg (single-colour body part) and another is for the left upper arm (multiple-colour body part). As shown in Figure 5.8 (a), the pixels specified by the contaminated correct estimations for the left lower leg are clustered using the approach described above. Two clusters are obtained where Cluster #1 represents the target pixels and Cluster #2 represents the background pixels. In Figure 5.8 (b), we can see that the target pixels in two different colours are separated into Cluster #1 and Cluster #2 respectively. The background pixels are gathered into Cluster #3.

5.3.2.2 Identifying Target Pixels

After all colour feature vectors in set Y_m are divided into subsets $\{Y_m^j\}_{j=1}^l$, our aim is to identify the positive subsets that representing the target pixels and the negative subsets representing the non-target pixels from these subsets. The distance that pixels in each cluster are to the central axis of the limb as discussed in Section 5.3.1.2 is utilized for this purpose, to distinguish the positive subsets from the negative subsets. Note that set Y_m can be viewed as a set of pixels because each colour feature vector in set Y_m corresponds to a pixel from the pixels specified by the *contaminated correct estimations* in set U_m .

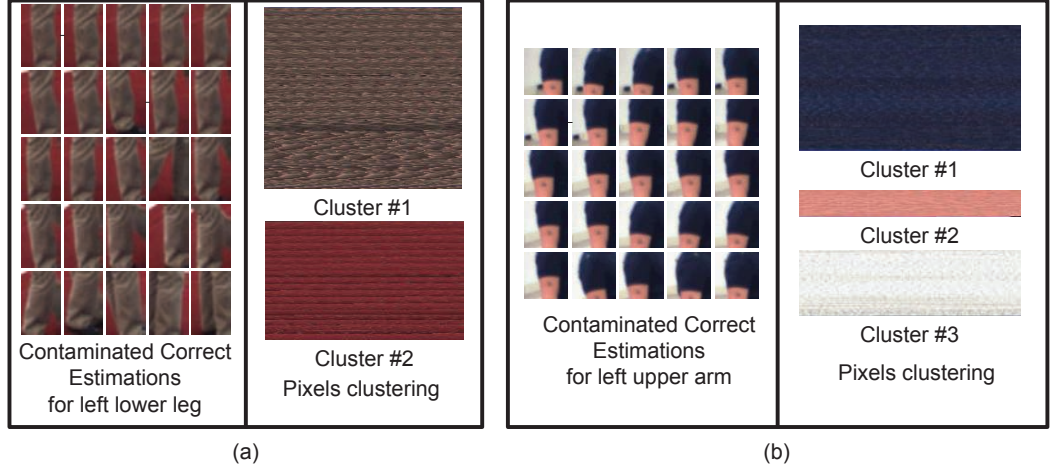


Figure 5.8: Two examples for pixels clustering. Figure (a) is the result of pixels clustering for left lower leg (single-colour body part). Figure (b) is the result of pixels clustering for left upper arm (multiple-colour body part).

Since the target pixels are more likely to appear in the central area of the bounding box while the non-target pixels are more likely to appear in the outer border, it is reasonable to assume that a subset Y_m^j in which the average distance of pixels with regard to the central axis is less than that of all pixels in Y_m can be identified as a positive subset. Otherwise it is considered a negative subset.

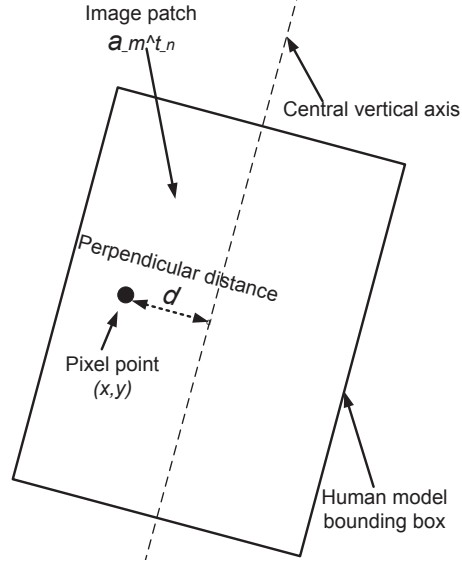


Figure 5.9: A demonstrating figure of image patch $a_m^{t_n}$ to explain how to compute a perpendicular distance d of the pixel (x, y) with respect to the vertical axis of the bounding box.

To identify the positive subsets from all the subsets, an average distance D_m^j for a subset Y_m^j is defined as follows. Any vector K_m^i in Y_m corresponds to a pixel (x, y) in certain image patch $a_m^{t_n}$ which is enclosed by a bounding box. As shown in Figure 5.9, in the image patch $a_m^{t_n}$, for a pixel of coordinate (x, y) , a perpendicular distance d with respect to the central vertical axis of the bounding box can be computed. Thus every K_m^i in Y_m would correspond to a perpendicular distance d_m^i . For a subset Y_m^j , an average distance D_m^j can be defined as

$$D_m^j = \frac{1}{N} \sum_{s=1}^N d_m^{i^s}, \quad (5.5)$$

where N is the size of set Y_m^j .

Beside the average distance for the pixels in each subset Y_m^j , the average distance D_m for all pixels in set Y_m is defined as

$$D_m = \frac{1}{\mathbb{N}} \sum_{t=1}^{\mathbb{N}} d_m^t, \quad (5.6)$$

where \mathbb{N} is the size of set Y_m .

If D_m^j is small, the corresponding pixels of Y_m^j are more likely to appear in the central area of the bounding box. According to the previous assumption that a subset Y_m^j where $D_m^j < D_m$ is identified as a positive subset, the potential positive subsets can be obtained by comparing their average distances with the average distance D_m . In order to avoid false positive subsets that are composed of noise pixels, any subset whose number of elements is far less than the others are removed. Based on this consideration, the threshold used to divide the positive subsets and false positive subset is set. In our experiment, given the number of pixels in all potential positive subsets and the number of potential positive subsets, which are respectively denoted as ω and ϕ , if the number of pixels in a potential positive subset is more than $\frac{2 \times \omega}{\phi}$, this subset is retained otherwise it is discarded. The positive subsets $\{Y_m^{j^r}\}_{r=1}^R$ are obtained by checking though all subsets. Each positive subset represents a type of target pixels.

5.4 Building Target-pixel Classifiers and Labelling Target Pixels

In the previous section, the target pixels (for body part m) from the pixels specified by the contaminated correct estimations in set U_m have been identified. Since set U_m consisting

of all the contaminated correct estimations for body part m is a subset of set S_m consisting of all preliminary estimations for body part m in a sequence, the identified target pixels for body part m are a part of all target pixels for the same body part in the sequence. In this section all target pixels (for body part m) in the sequence are to be identified using the known target pixels. To achieve this aim, target-pixel classifiers for body part m will be learned using the known target pixels from the contaminated correct estimations in set U_m . The target-pixel classifiers are then applied to identify all target pixels (for body part m) in each frame of a sequence.

5.4.1 Building Target-pixel Classifiers

In order to identify and label all target pixels in each frame, target-pixel classifiers are expected to be learned when the positive subsets containing the target pixels and the negative subsets containing the non-target pixels from the contaminated correct estimations are available. In our system, a set of simple Gaussian classifiers are learned using the pixels in the positive subsets $\{Y_m^{j^r}\}_{r=1}^R$. Although a more complex classifier, such as SVM or a quadratic logistic regression classifier, can be learned due to the availability of both the target pixels and non-target pixels, it is found to be unnecessary because a set of Gaussian classifiers proves to be sufficient in our human tracking system.

As discussed before, if there is only one positive subset in $\{Y_m^{j^r}\}_{r=1}^R$ (resulting from a single-colour body part), only one Gaussian classifier will be learned for body part m . If there are multiple positive subsets in $\{Y_m^{j^r}\}_{r=1}^R$ for a multiple-colour body part, a set of Gaussian classifiers will be learned for body part m . The number of Gaussian classifiers for a body part is determined by the number of the positive subsets. Regardless of the number of Gaussian classifiers for a body part, each Gaussian classifier is learned using the following method.

Given the data in $Y_m^{j^r}$, the Gaussian parameters including the mean $\mu_m^{j^r}$ and the covariance matrix $\Sigma_m^{j^r}$ can be estimated by maximum likelihood estimation. To classify an unknown vector x , the likelihood of an unknown vector is defined as

$$p(x|Y_m^{j^r}) = N(x, \mu_m^{j^r}, \Sigma_m^{j^r}). \quad (5.7)$$

If $p(x|Y_m^{j^r})$ is more than γ , the vector x is identified as belonging to the target pixels that is represented by the positive subset $Y_m^{j^r}$. A threshold γ is set to 0.08 in our experiment.

5.4.2 Applying Target-pixel Classifiers to Each Frame

Due to the fact that a contaminated correct estimation cannot perfectly cover all target pixels of its corresponding body part, there is a possibility that part of the target pixels for that body part are not identified. In addition, there might be frames for which correct estimations (even contaminated) are not available since all their preliminary estimations were identified as incorrect. We attempt to build specific appearance models that can also be applicable for these frames, hence the target pixels locate in such frames need also be identified. Our goal is to build target-pixel classifiers that can be applied for every frame in a sequence.

Each body part corresponds to a set of target-pixel classifiers and thus each classifier for a body part can be applied to check each pixel in every frame of a sequence. Consequently, for each classifier of a body part, a binary image (called *a mask image*) can be generated for every frame in a sequence. The mask image for every frame has identical size to the image of the frame, and functions as a mask indicating which pixels are classified as the target pixels corresponding to the body part. If a pixel in a frame with coordinate (x, y) is a target pixel, a corresponding pixel in the mask image with coordinate (x, y) is marked as 1, otherwise marked as 0.

An experiment is conducted in the Combo sequence from HumanEva (as shown in Figure 5.10 (a)) to test the proposed target-pixels classifiers. In this experiment, the learned target-pixel classifiers for torso, leg, upper arm, lower arm and head are respectively used to mark their target pixels. The single-coloured body part such as torso, leg and lower arm etc. corresponds to only 1 target-pixel classifier and thus only one type of target pixels in these body part need to be identified, as shown in Figure 5.10. In contrast, the two-coloured (multi-coloured) body part such as the upper arm and head has two target-pixel classifiers and thus two types of target pixels for a body part are separately identified, as shown in Figure 5.11. These mask images will subsequently be used as the limb detectors for a second, appearance-based pass of human pose estimations.

Single-coloured body part It can be seen from Figure 5.10 (a) that the colour of torso in this sequence is significantly different from the colours of other body parts, so the torso pixels are effectively identified as shown in Figure 5.10 (b). Although a few noise pixels are marked as the torso pixels, they have a minimal impact on localizing the torso. Unlike the torso with a unique colour, due to the symmetrical structure of a human body, the symmetrical body parts such as left and right arms/legs usually have the same colour. For example, Figure 5.10 (a) shows that the legs including left/right lower/upper legs have



Figure 5.10: Several examples of marking target pixels for single-coloured body part using learned Gaussian appearance classifiers. The frames shown in this figure are representative and typical in Combo sequence (Human_Eva_I_S2_Combo_2_C2) from HumanEva dataset (Sigal and Black, 2006).

the same colour, so four different target-pixel classifiers will represent the same colour. No matter which target-pixel classifier is used, the target pixels for all the four leg parts are identified and marked, as shown in Figure 5.10 (c). It is impossible to decide which leg part these marked pixels come from based only on the target-pixel classifiers. In addition, sometimes some asymmetrical body parts are also of the same colour. For example, as shown in Figure 5.10 (a), the face has the same colour with the lower arm. When using the lower arm classifier or head classifiers to mark the target pixels, pixels from both the lower arm and the head will be marked as shown in Figure 5.10 (d). Fortunately, the pictorial structure can be used to resolve the confusion between body parts. The spatial relations between body parts are defined in the pictorial structure, which helps to determine which target pixels should belong to which body part.

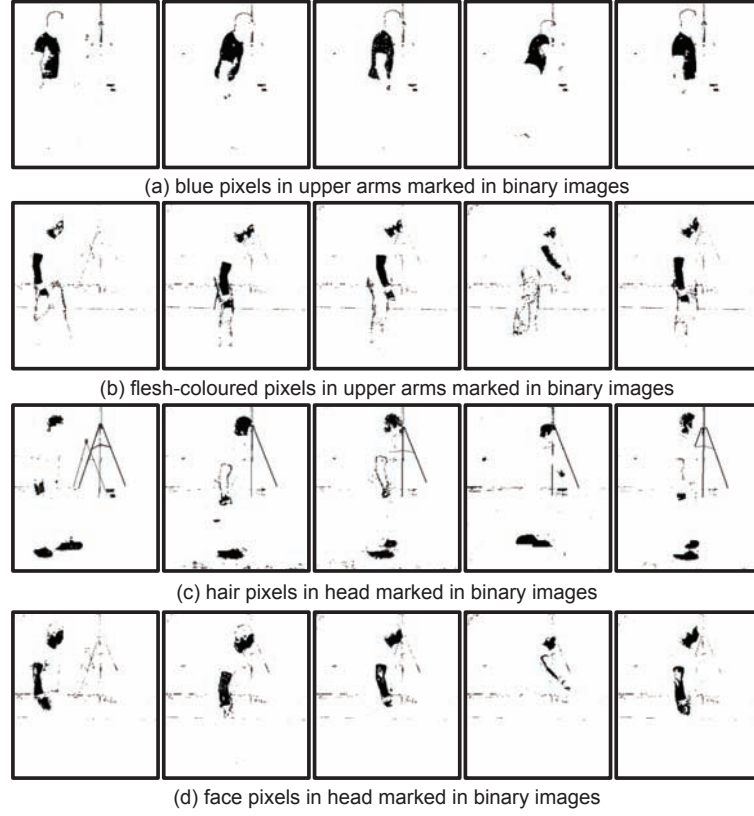


Figure 5.11: Several examples of marking target pixels for multiple-coloured body part using learned Gaussian appearance classifiers.

Multi-coloured body part Unlike the single target-pixel classifier for a single-coloured body part such as the torso, upper arms and lower/upper legs in the Combo sequence, multiple target-pixel classifiers are obtained for every multi-coloured body part such as the head and upper arms in the same sequence. Specifically, two target-pixel classifiers are built for the head or the upper arms in this sequence and each target-pixel classifier can be used to identify a type of target pixels in the body part. For example, there are two types of target pixels in the upper arms, which are blue pixels and flesh-coloured pixels. Correspondingly, two different target-pixel classifiers are obtained by the proposed approach. After applying them separately to mark their corresponding target pixels, the results are shown in Figure 5.11 (a) and (b). In the same way, the two types of target pixels for the head are marked as shown in Figure 5.11 (c) and (d).

5.5 The Uncontaminated Appearance Model for Tracking

After the target pixels and non-target pixels for each body part have been marked for each frame of a sequence, the specific appearance features for each body part are recorded in these mask images. They can be utilized to estimate human pose in each frame. Similar to the approach proposed in Chapter 4, generic detections are obtained by utilizing the specific appearance of the human body thus obtaining the candidates that satisfy both the generic and specific appearances of the human body. The specific appearance feature used here is recorded in the mask images, which is different to the colour histogram feature used in Chapter 4. In order to examine the specific appearance feature of a generic detection based on these markers, the appearance template for each body part must first be defined. The response of the specific appearance for a generic detection is generated by matching its corresponding appearance template with the markers for the target pixels and non-target pixels. The final estimation is determined by choosing the candidate with the best specific appearance response.

5.5.1 Specific Appearance Template

Each body part in the human model corresponds to a specific appearance template, as shown in Figure 5.12 (a). The purpose of this template is to verify that a detection is matching the body part (target) pixels in the expected location of target pixels (ie: centrally), thereby ensuring the detection is well-oriented on the part. The size of a specific appearance template for a body part is the same as the size of its bounding box. A specific appearance template is composed of two parts: target-pixel area and non-target-pixel area. The centre of the target pixel area locates in the centre of the template. The non-target pixel area that occurs in the border of a template is around the target pixel area. If a target-pixel marker appears in the target-pixel area, a positive response is obtained, otherwise a negative response is obtained. The same applies to the non-target pixel marker.

The template for an individual body part is shown in Figure 5.12 (b). The green area is the target-pixel area and the red area is the non-target-pixel area. The width and length of the template for body part m are denoted as w_m^o and l_m^o , while the width and length of the green area (target-pixel-area) are denoted as w_m^p and l_m^p respectively. Given the width w_m and length l_m of the bounding box for body part m , the four parameters $(w_m^o, l_m^o, w_m^p, l_m^p)$ of a template for a single-coloured body part can be determined, i.e., $w_m^o = w_m$, $l_m^o = l_m$, $w_m^p = w_m - 4$ and $l_m^p = l_m - 4$.

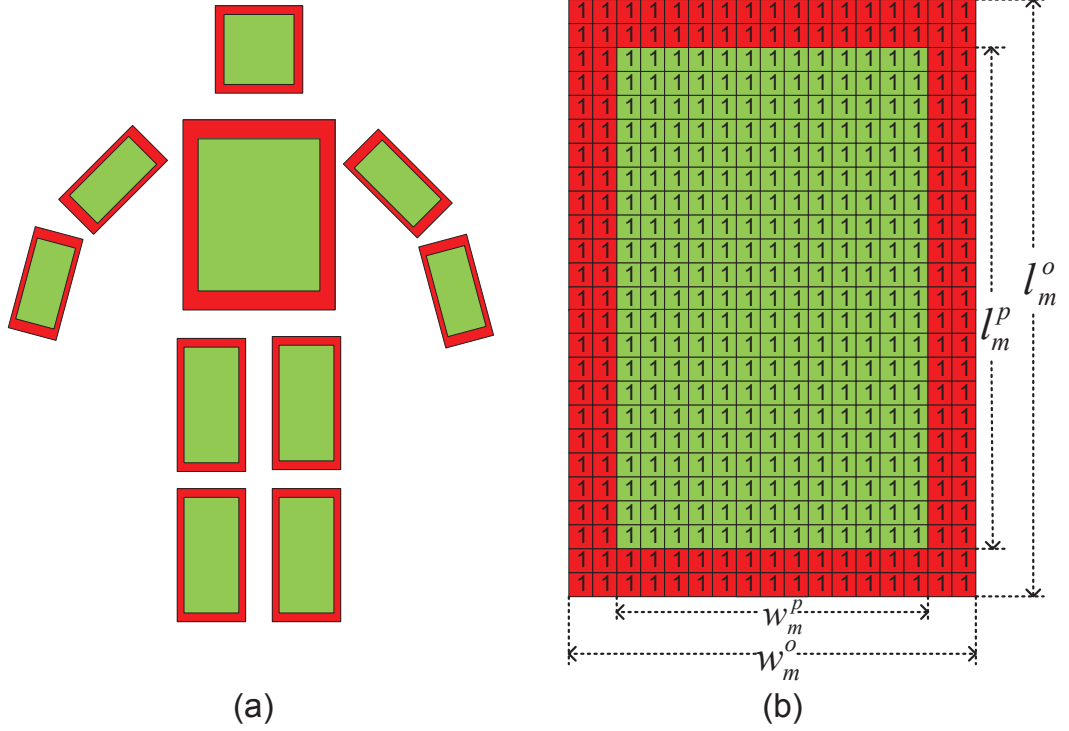


Figure 5.12: (a) Each body part in the human model relates to a specific appearance template where green represents the target-pixel area and red represents the non-target-pixel area. (b) Appearance template for an individual body part.

5.5.2 Appearance Likelihood Model

After the specific appearance template for each body part is defined, it can be used to examine the specific appearance of a generic detection. To achieve this aim, an appearance likelihood model for each body part will be built.

When examining the specific appearance of a generic detection e_m^t for body part m in frame t , given the width and length of its bounding box w_m and l_m and a set of mask images $\{b_k\}_{k=1}^N$ for body part m in frame t where N is the number of types of target pixels in body part m , a set of mask patches $\{p_k\}_{k=1}^N$ that corresponds to the generic detection e_m^t can be extracted from the set of mask images $\{b_k\}_{k=1}^N$. The width w_m^o and the length l_m^o for the appearance template of body part m is set as w_m and l_m , while the width w_m^p and the length l_m^p of the green area (target pixels area) are set as $w_m - 4$ and $l_m - 4$, as shown in Figure 5.12 (b). When using the specific appearance template of body part m to match the set of mask patches $\{p_k\}_{k=1}^N$, the response $r(i, j)$ at coordinate (i, j) of the

template is defined as:

$$r(i, j) = \begin{cases} 1 & \text{if } \prod_{k=1}^N p_k(i, j) = 0 \text{ and } a_m(i, j) = 1 \\ 1 & \text{if } \prod_{k=1}^N p_k(i, j) = 1 \text{ and } a_m(i, j) = 0 \\ 0 & \text{otherwise} \end{cases} \quad (5.8)$$

where $1 \leq i \leq w_m$, $1 \leq j \leq l_m$ and the function $a_m(i, j)$ is defined as

$$a_m(i, j) = \begin{cases} 1 & \text{if coordinate } (i, j) \text{ locates in the green area of this template} \\ 0 & \text{if coordinate } (i, j) \text{ locates in the red area of this template} \end{cases} \quad (5.9)$$

Finally, the appearance likelihood of the generic detection $L_m(e_m^t)$ is defined as

$$L_m(e_m^t) = \frac{\sum_{i=1}^{w_m} \sum_{j=1}^{l_m} r(i, j)}{w_m \times l_m} \quad (5.10)$$

5.5.3 Tracking

After the appearance likelihood model for each body part is built as described, a second pass of tracking can be performed by utilizing the specific appearance feature to achieve a more accurate estimation. The aim of this second pass is to find the best estimation from the generic detections that satisfy both the generic and specific appearance features. Candidates that satisfy both the generic and specific appearance features are first identified by filtering through all generic detections. Spatial search is then applied to find the best estimation among the candidates.

5.5.3.1 Filtering Generic Detections

As described in Chapter 3, two types of generic detections including the preliminary estimation and multiple alternative estimations are derived from the generic pose detector. For any frame, the preliminary estimation for body part m is the optimal (posterior) estimation, which is the best match with the generic appearance model under the pictorial structure. Beside the optimal estimation, a set of sub-optimal estimations (i.e., multiple alternative estimations) for body part m , denoted as U_m , is obtained by sampling the top N posterior except for the maximum posterior (where $N = 30$ for the experiments in Section 3.4.2). The estimations in this set also match the generic appearance model under the pictorial structure, but not as well as the preliminary estimation. Among these top generic

estimations, candidates that satisfy both the generic and specific appearance models need to be found. They can be generated by filtering through these generic detections using the specific appearance likelihood model.

As discussed in Section 5.5.2, the appearance likelihood of a generic detection e_m^t for body part m in frame t is denoted as $L_m(e_m^t)$. Given any estimation u in a set G_m of the top generic detections for body part m in frame t , if $L_m(u) > 50\%$, it is retained otherwise it is discarded thus generating a subset $G_m^* \subseteq G_m$ in which each element satisfies both the generic and specific appearance models. The subset G_m^* is called the concentrated set of top few generic detections and each estimation in G_m^* is a candidate for the final estimation.

5.5.3.2 Spatial Search

After the concentrated set of top generic detections G_m^* is obtained, it is necessary to determine which candidate in this set should be chosen as the final estimation. Spatial search is used to obtain the final estimation. In the spatial search, except the torso which is the root of the body structure tree, the configuration of each body part is restricted by the configuration of its parent. Details are described in Section 4.3.2.2. The final estimation of torso is determined by choosing the highest-likelihood candidate in the concentrated set.

5.6 Experiments and Discussion

In order to evaluate the performance of the proposed uncontaminated specific appearance model for tracking, it is applied to three challenging sequences. The experimental results are also compared against the results of three other systems: Ramanan system, Ferrari-Core system and the tracking system proposed in Chapter 4. For the sake of simplicity and convenience, the tracking system proposed in this chapter is called the CLASSIFIER system, while the tracking system proposed in Chapter 4 is called the CLUSTERING system.

Three sequences, as shown in Figure 5.13, are used to test the four systems. One is the ‘Baseball’ dataset of Ramanan *et al.* (2007) and two are from the HumanEva data set (Sigal and Black, 2006), ‘HE1_S2_Walking_1.C1’, and ‘HE1_S2_Combo_2.C2’, which are simply called Walking and Combo. The Baseball video is a sequence of 200 frames used by Ramanan *et al.* (2007) that records a pitcher throwing out a ball. The Walking

sequence and Combo sequence record the same person wearing identical clothing but acting different motions. Moreover, the two sequences are recorded from different camera angles. The Walking sequence includes 157 frames (from frame 353 to frame 509 in the video) and the Combo sequence includes 400 frames (from frame 1855 to frame 2254). They are used to demonstrate that our system can be widely applied to many different motions, with different viewing angles and different numbers of frames. The ground truths in these sequences are manually marked for evaluating the performance of the various systems. The same evaluation metrics as described in Chapter 4 are used here to evaluate the in tracking performances of the four systems.

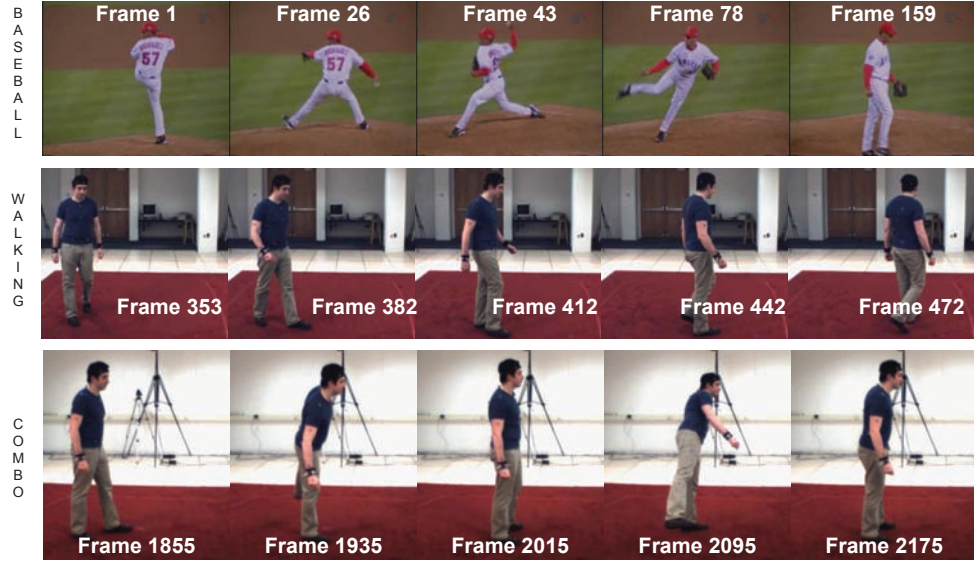


Figure 5.13: Three sequences are used to compare our proposed system with three other systems (ie: Ramanan’s system, Ferrari-Core system, and CLUSTERING system). Walking is short for HE1_S2_Walking_1_C1. Combo is short for HE1_S2_Combo_2_C2.

5.6.1 Baseball and Walking Sequences

Experiments are first conducted to test the CLASSIFIER system on the Baseball sequence and Walking sequence. Figure 5.15 gives the screenshot of tracking results from CLUSTER and CLASSIFIER, which shows the difference of tracking results in the two systems. Figure 5.14 and 5.16 show the comparison of the CLASSIFIER system against three other systems in tracking performance based on Metric-1 and Metric-2.

According to the evaluation results based on Metric-1 in Figure 5.14, the CLASSIFIER system obviously outperforms the other three systems although it does not achieve significant improvement in comparison to the CLUSTER system. Since a detailed comparison

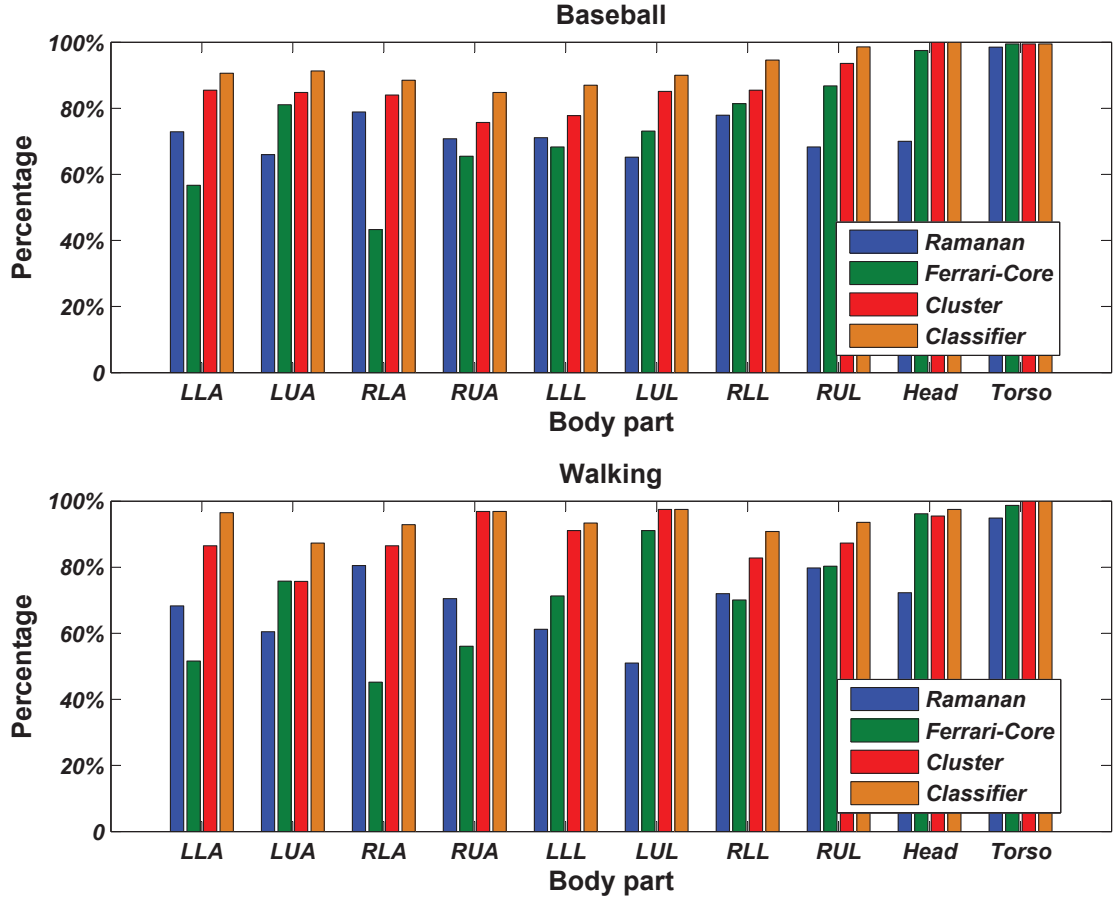


Figure 5.14: The performance comparison between Clustering and Classifier systems based on Metric-1 in Baseball and Walking Sequences.

of the CLUSTER system to Ramanan system and Ferrari-Core system has been presented in Chapter 4, which has demonstrated that CLUSTER outperforms the other two systems, here we focus mainly on the comparison between the CLUSTER and CLASSIFIER systems. For the Baseball sequence, both systems achieve similar performance in tracking torso and head. The performance of CLASSIFIER is slightly better in tracking the remaining body parts, but the difference is not significant. Specifically, the tracking performance for LLA, LUA, RLA, RUA is improved respectively by 5.1%, 6.5%, 4.5% and 9.1%. The tracking performance for LLL, LUL, RLL and RUL is improved respectively by 9.2%, 4.9%, 9.1% and 5%. For the Walking sequence, the performance for both system is also comparative overall, although CLASSIFIER is slightly better in tracking some body parts such as LLA (improved by 10%), LUA (improved by 11.6%), RLA (improved by 6.4%), RLL (improved by 8%) and RUL (improved by 6.3%).

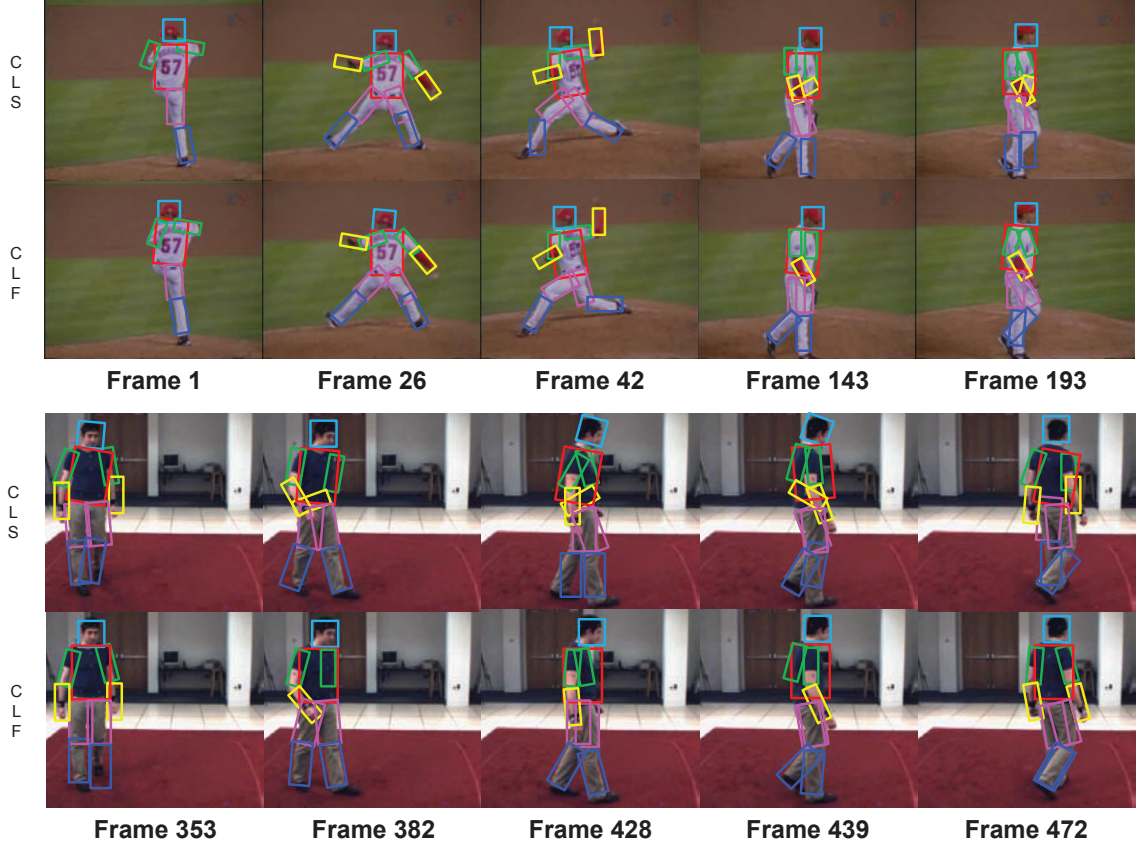


Figure 5.15: The screenshot of tracking results in Baseball and Walking sequences from CLUSTER (CLS) and CLASSIFIER (CLF).

Although small, there is still a more than 5% improvement by the CLASSIFIER system. The improvement arises from the fact that the CLASSIFIER system learns a more accurate appearance model than CLUSTER. The essential difference between the CLUSTER and CLASSIFIER is the different ways in which they learn their specific appearance models. As discussed in Section at the beginning of the chapter, when learning the specific appearance model in CLUSTER, non-target pixels are unavoidably included in the training samples and thus the learned specific appearance model is contaminated. When the contaminated specific appearance model is used to examine the specific appearance of generic detections, although most negative generic detections can be identified and removed, there are still some negative generic detections possibly accepted by the contaminated specific appearance model. For example, the estimations for lower arms in Frame 143 and 193 of the Baseball sequence and Frame 382, 428 and 439 of the Walking sequence are shown in Figure 5.15. In these instances, the generic detections all consist of a certain amount of target pixels, which make the specific appearance of them reach a point that can be

accepted by the contaminated specific appearance model. In addition, the contaminated specific appearance model also results in situations that the estimations are not accurate enough (ie: include too many non-target pixels) even though they are considered as correct estimations under Metric-1, e.g., estimations for legs in Frame 42, 193 of the Baseball sequence and Frame 382, 428 of the Walking sequence. In contrast to CLUSTER, CLASSIFIER learns a specific appearance that are not (or less) contaminated by the non-target pixels and thus the learned specific appearance model can more clearly distinguish the target pixels from the non-target pixels, resulting in more accurate estimations, as shown in Figure 5.15.

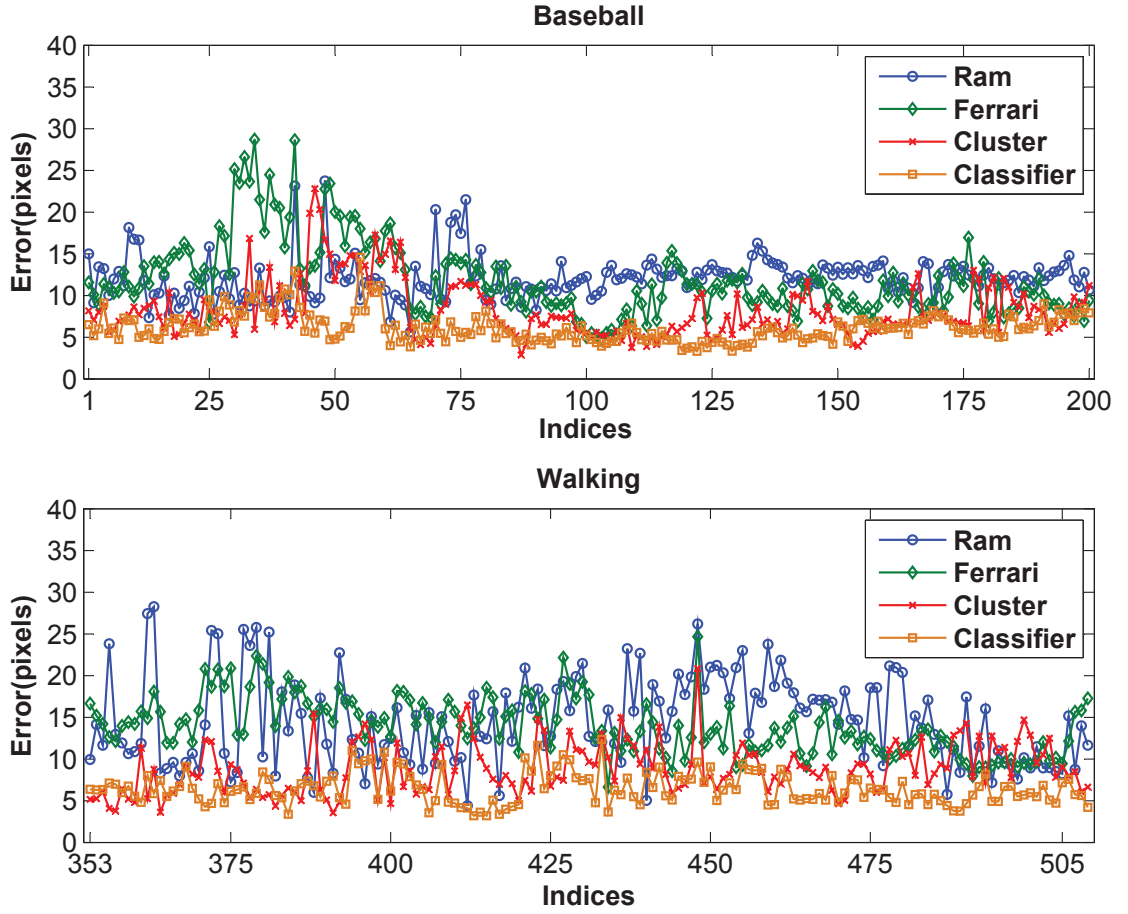


Figure 5.16: The performance comparison between Clustering and Classifier systems based on Metric-2 in Baseball, Walking, Throwing and Combo sequences.

The advantage of the CLASSIFIER system can be further demonstrated by the evaluation results based on Metric-2. As shown in Figure 5.16, it can be clearly seen that the CLASSIFIER system achieves noticeably more accurate estimations than the other three systems (including CLUSTER. For the Baseball sequence, CLASSIFIER obtains more accurate

estimations than Ramanan’s system in 195 (97.5%) of all the 200 frames, than Ferrari-Core system in 196 (98%) of all 200 frames, and than the CLUSTER system in 148 (74%) of all 200 frames. The average error (in terms of pixel distance) for CLASSIFIER is 6.1, which is less than 8.2 for CLUSTER, 12.0 for Ramanan’s system and 11.8 for Ferrari-Core system. For the Walking sequence, CLASSIFIER achieves more accurate estimations than Ramanan’s system in 153 (97.4%) of all 157 frames, than Ferrari-Core system in 156 (99.3%) of all 157 frames and than CLUSTER in 117 (74.5%) of all 157 frames. The average error (in terms of pixel distance) for CLASSIFIER is 6.5, which is less than 8.8 for CLUSTER, 14.5 for Ramanan’s system, and 13.9 for Ferrari-Core system.

5.6.2 Combo Sequence

The Combo sequence is used to further evaluate the performance of the CLASSIFIER system, which is also compared against the three other systems. Unlike the evaluation based on the Baseball and Walking sequence which is focused on the overall performance, this section evaluates the performance of CLASSIFIER by focusing on the performance in tracking individual body parts. On the basis of Metric-2, instead of measuring the average error of all joint points in a human, the performance of a system in tracking each body part is evaluated by measuring the error of two end points (joint points) of each body part. The joint points that are irrelevant to the body parts in question are ignored.

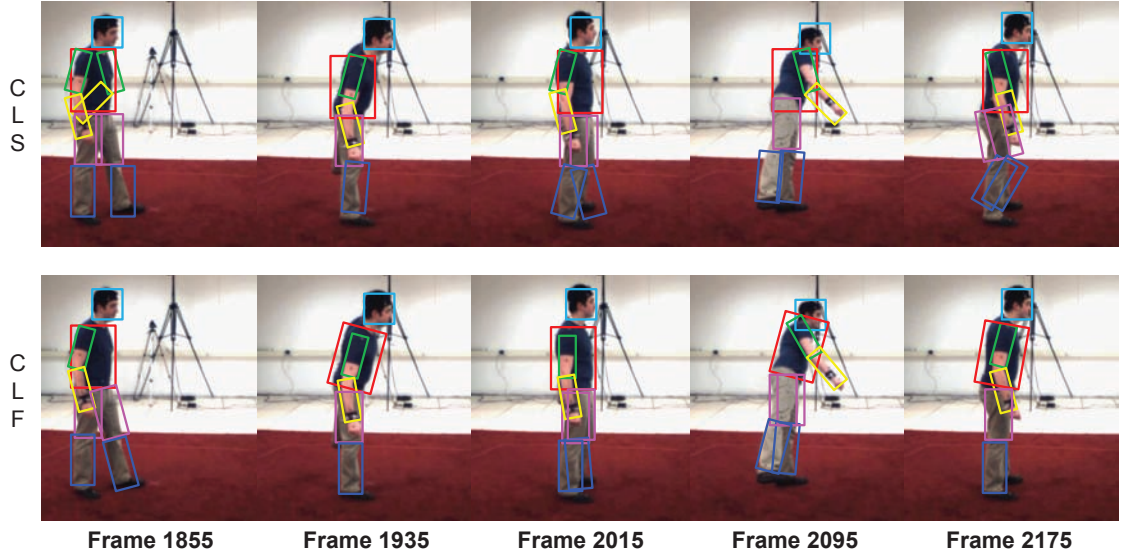


Figure 5.17: The screenshot of tracking results in Combo sequences from CLUSTER (CLS) and CLASSIFIER (CLF).

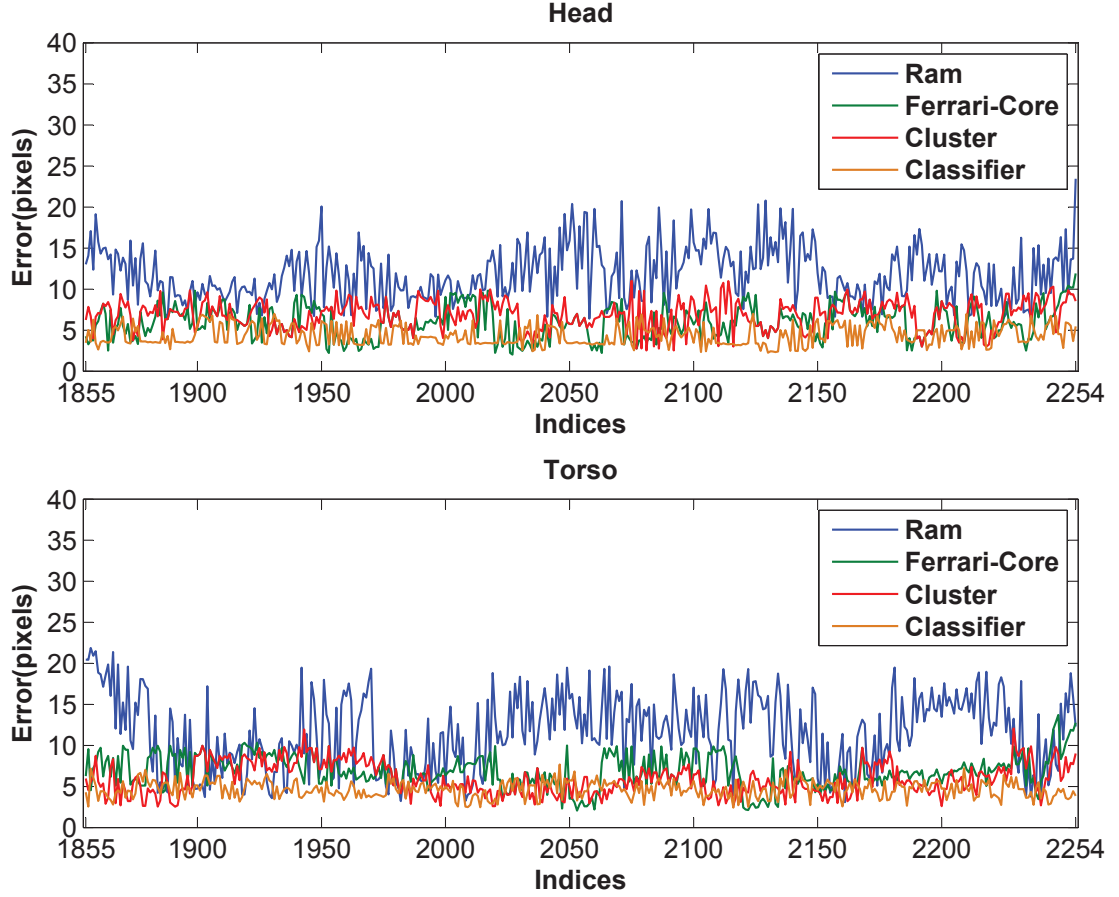


Figure 5.18: The errors in estimating the configuration of head and torso in every frame of Combo sequence by four systems.

The performance of the four systems in tracking head and torso is given in Figure 5.18. It can be clearly seen that three systems, namely, Ferrari-Core system, CLUSTER and CLASSIFIER, achieve similar performance in tracking head and torso but obviously outperform Ramanan’s system. More specifically, for head, the average errors from Ferrari-Core system, CLUSTER and CLASSIFIER are respectively 6.8, 5.9 and 4.3, which are much less than 11.8 from Ramanan’s system. For torso, the average errors from Ferrari-Core, CLUSTER and CLASSIFIER are respectively 6.8, 5.9 and 4.6, which are also far less than 11.5 from Ramanan’s system. This is mainly due to the fact that the generic (shape) appearance model is used in Ferrari-Core system, CLUSTER and CLASSIFIER. but is not used in Ramanan’s system. As analyzed in Section 4.4.2.1, a system based on only the specific (colour) appearance model (such as Ramanan’s system) is more likely to cause confusion in tracking head and torso than a system using both the generic (shape) and specific (colour) appearance models. In fact, in comparison with other body parts, the shape features of torso

and head are more easily identifiable. Specifically, the head is a squarish shape which is different from the rectangular shape of the torso, legs or arms. The torso has the biggest size among all body parts and it is in a rectangular shape similar to the legs or arms. Moreover, they are almost never (fully) occluded by other body parts, which also makes them easier to be detected. Based on the above facts, a generic (shape) appearance model is usually effective and accurate enough in estimating the configuration of the type of body parts which have a distinctive shape feature.

It is worth noting that the accuracy of the CLASSIFIER system is still slightly better than Ferrari-Core and CLUSTER in tracking head and torso. This means that a better specific (colour) appearance model still generate more accurate estimations even when a generic (shape) appearance model plays a decisive role in tracking a body part with a distinctive shape feature.

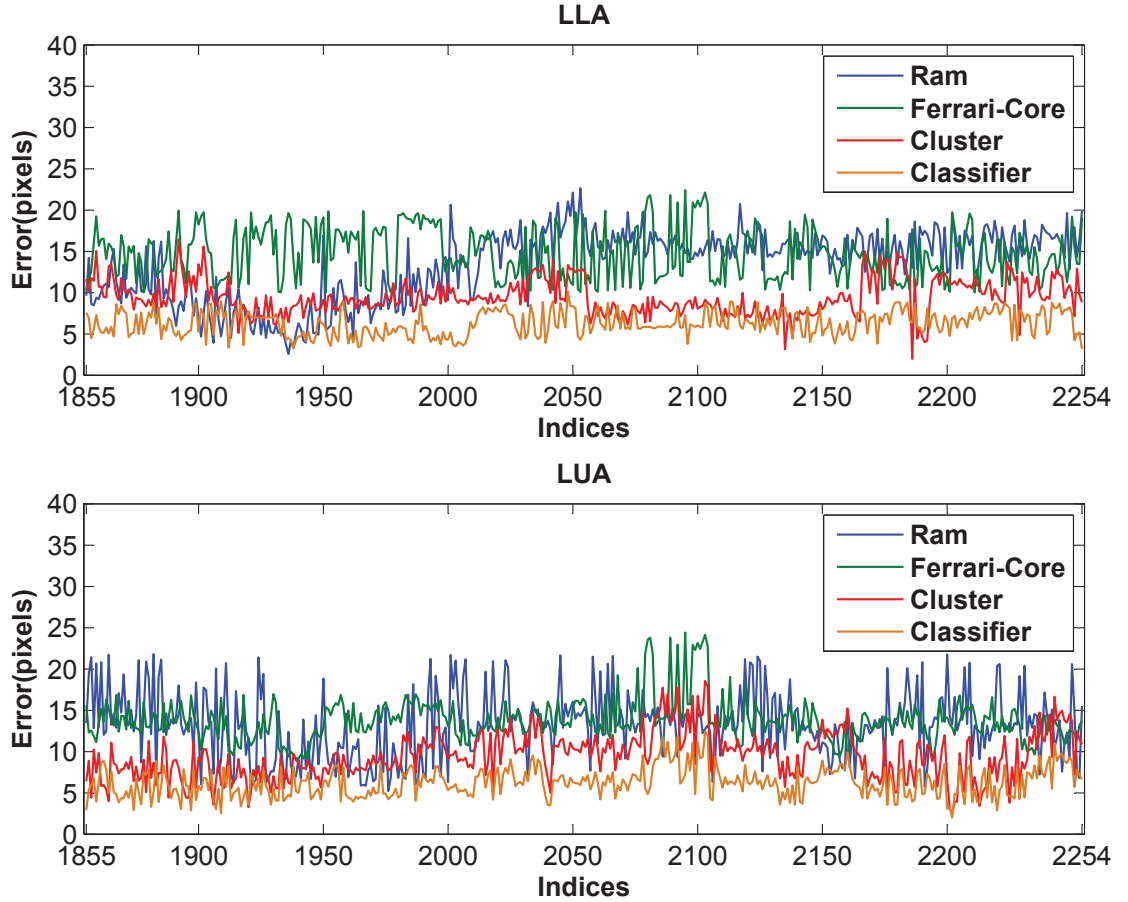


Figure 5.19: The errors in estimating the configuration of LLA and LUA in every frame of Combo sequence by four systems.

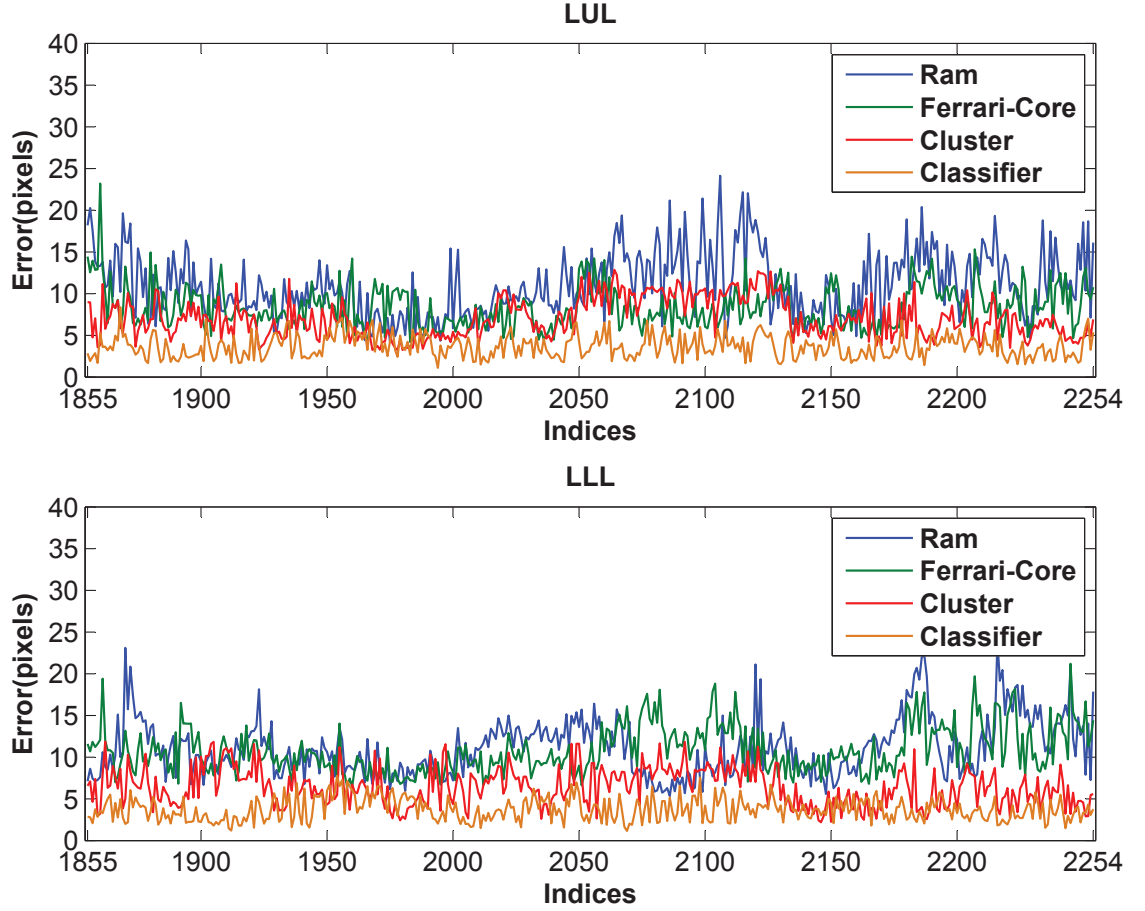


Figure 5.20: The errors in estimating the configuration of LUL and LLL in every frame of Combo sequence by four systems.

As commonly acknowledged, the most challenging body parts to be tracked are arms. There are two reasons here: 1) in comparison with torso and leg the size of an arm is smaller; 2) the arms are much more flexible than torso and legs in motion. Therefore, the performance in tracking arms can best represent the performance of a tracking system. Figure 5.19 shows the performance of the four systems in tracking LLA and LUA. We can see that the tracking accuracy is significantly improved by the CLASSIFIER system in comparison with Ramanan's system, Ferrari-Core system and the CLUSTER system. For LLA, the average errors from Ramanan's system, Ferrari-Core system and CLUSTER are 13.3, 14.7 and 9.4 respectively, whereas the average error from CLASSIFIER is 6.3. For LUA, the average errors from Ramanan's system, Ferrari-Core system and CLUSTER are respectively 15.8, 14.3 and 9.1, whereas the average error from CLASSIFIER is 6.6. As discussed in Chapter 4, CLUSTER outperforms Ramanan's system and Ferrari-Core system in tracking arms. Here we analyze why CLASSIFIER achieves a better accuracy

than CLUSTER. In contrast to the torso and head, the arms do not have a distinctive shape feature. The accuracy of tracking them depends to a great extent on the accuracy of the specific (colour) appearance model. The CLASSIFIER system that is built on an uncontaminated (or less contaminated) specific appearance model can achieve much better performance than the CLUSTER system in which the specific appearance is contaminated by non-target pixels.

Besides tracking arms, the effectiveness of the uncontaminated specific appearance model is also demonstrated in tracking legs, as shown in Figure 5.20. The effectiveness of the proposed approach for tracking legs is for similar reasons as for the arms.

5.7 Chapter Summary

Chapter 4 proposed a tracking system where a specific appearance model is based on the colour histogram features of the preliminary estimations. In this process, non-target pixels are possibly included in the extracted features, so the consequent learned appearance model is contaminated and the accuracy of the learned appearance model is compromised.

In order to learn an uncontaminated specific appearance model, which is not contaminated by non-target pixels, this chapter first presents an algorithm to identify and remove non-target pixels from preliminary estimations. Specifically, the characteristics that can distinguish target pixels with non-target pixels are investigated and analyzed. It has been found that in the bounding box specified by a preliminary estimation, target pixels are more likely to appear on the central area than the outer border. Based on these characteristics and pixel clustering, target pixels and non-target pixels in preliminary estimations can be separated.

Next, the algorithm for building an appearance likelihood model for each body part using the obtained target pixels is presented. Specifically, after target pixels for each body part are identified and extracted, they are used to learn a set of target-pixel classifiers for each body part. Target pixels for each body part in each frame of a sequence can be marked by these target-pixel classifiers. An appearance template for each body part is then defined according to its shape feature. The appearance likelihood for a body part can be obtained by matching the target pixels through the appearance template for that body part.

Finally, the final estimation for each body part in each frame is derived by choosing the best estimation using the uncontaminated specific appearance feature from the generic

detections. Experiments are conducted to compare the proposed CLASSIFIER system with the CLUSTER system proposed in Chapter 4, Ramanan's system and Ferrari-Core system. Results show that the CLASSIFIER system obviously outperforms the CLUSTER system, as well as Ramanan's system and Ferrari-Core system.

Chapter 6

Conclusions

This thesis has proposed a system based on some novel algorithms for estimating 2D human configurations over time based on monocular view. There is no restriction on the human activities and the camera parameters are unknown. In this way, it addresses the same problem domain as two successful systems recently proposed in the literature ([Ramanan *et al.*, 2007](#); [Ferrari *et al.*, 2009](#)). The two existing approaches achieved good performance but there are some weaknesses in their formulation that this thesis has analyzed and proposed alternative approaches to overcome.

In Chapter 3, the framework of the pictorial structure ([Felzenszwalb and Huttenlocher, 2005](#)) and a corresponding generic human pose detector are described. In order to test the assumptions of [Ferrari *et al.* \(2009\)](#), the relationship between the accuracy and the likelihood of the optimal estimations between different frames were explored. The results show that there is no useful relationship between accuracy and likelihood for comparing optimal estimations between frames. Subsequently, the generic human pose detector is evaluated and analyzed by testing against two representative sequences. First, the percentage of frames where the optimal estimations are correct (in the sense of a 50% or more overlap with the ground-truth) is computed for each body part. Results show that the correct detection rate of the optimal estimations for most body parts are between 50%-70% (except for the prominent torso and head whose detection rate is over 80%). This means that a small majority of the optimal estimations are accurate, but the overall set of the optimal estimations cannot be straightforwardly relied on.

Therefore, in order to learn an accurate specific appearance model based on the set of the optimal estimations, it is necessary to define a method to carefully select the accurate estimations in this set. Fortunately, since the correct detection rate of the optimal estimations for each body part does exceed 50%, it provides an opportunity to distinguish correct optimal estimations from incorrect optimal estimations by clustering the optimal estimations according to the similarity of colour appearance. Under the assumption that the correct optimal estimations will have similar colour appearance and the correct rate of the optimal estimations for each body part is in the majority (or at least the largest single group), the largest cluster will represent the correct optimal estimations for building

a specific appearance model. However, to find correct detections in all frames rather than just those with correct optimal estimations, this thesis proposes to search for good colour matches within the top N estimations. To verify that this can reasonably lead to good body part detections, an analysis was conducted to investigate the percentage of frames where correct estimations exist in the top N most likely estimations for each body part. The results show that the correct estimation is found within the top 30 most likely estimations in 90% of frames, indicating that it is feasible to find a correct estimation that conforms to both generic and specific appearance model by filtering the top 30 most likely estimations using a specific appearance model.

In Chapter 4, a tracking system for estimating 2D human pose over time is implemented based on the findings of Chapter 3. In this system, two important approaches are proposed:

1. To learn a specific appearance model using clustering.
2. To obtain final estimations that satisfy both the generic and specific appearance models.

A specific (colour) appearance model is learned based on the generic detections of the generic detector. As with [Ramanan *et al.* \(2007\)](#) and [Ferrari *et al.* \(2009\)](#), in order to learn an accurate specific appearance model, this thesis also automatically selects the accurate estimations from the optimal estimations of the generic detector. However, unlike [Ramanan *et al.* \(2007\)](#) and [Ferrari *et al.* \(2009\)](#) that assume the likelihood can infer the accuracy of estimations (an assumption shown to be problematic in Chapter 3), the continuity of human appearance in a sequence is utilized to split the optimal estimations into ‘accurate’ and ‘inaccurate’ estimations. To achieve this, colour histogram clustering is first applied to the optimal estimations to group together estimations with similar colour appearance. The largest cluster is then selected as the representative of the accurate estimations, based on the findings from Chapter 3 that generic estimations are accurate in most cases, if not the overwhelming majority. The estimations in this group are then chosen as the training samples to learn the specific appearance model. Experimental analysis demonstrates that the clustering method provides a significantly improved specific (colour) appearance model for tracking in comparison with [Ferrari *et al.* \(2009\)](#).

The chapter also proposed a novel approach to fusing evidence from generic and specific appearance models such that the final estimations will satisfy both models, unlike [Ramanan *et al.* \(2007\)](#) who discards evidence from the generic detections after building the specific model. Specifically, the final estimations in the proposed system are obtained by filtering multiple alternative estimations (ie: the top N most likely generic detections)

by the learned specific appearance model. This leverages the results of Chapter 3 which demonstrated that correct estimations are nearly certain to exist within the top 30 detections of a given body part. This is in contrast to Ferrari *et al.* (2009), who generates final estimations based on averaging the likelihoods of the generic appearance and a specific (colour) appearance models and thus does not ensure that both generic and specific evidence are highly likely to be satisfied at the same time.

The tracking system proposed is quantitatively tested and evaluated in comparison with Ramanan’s system and the core of Ferrari’s approach on several video sequences containing humans performing a variety of motions. Evaluation consists of two metrics, one based on bounding box overlap with the ground truth and the other metric of pixel error of joint locations. The bounding box metric labels an estimation as correct if the estimation covers at least 50% pixels specified by the ground-truth. The second evaluation metric is in terms of the pixel error distance of joint points between estimation and ground truth. The experimental results show that the proposed system outperforms those of both Ramanan *et al.* (2007) and Ferrari *et al.* (2009), often significantly. Specifically, the performance in bounding box overlap for tracking most body parts (except head and torso) is significantly improved by proposed system. Notably, the proposed system outperforms the existing systems significantly in tracking the more difficult extremities such as the arms and legs. The average accuracy of the proposed system in tracking arms is 85%, significantly exceeding that of Ramanan (71%) and Ferrari (65%). For legs, the proposed system achieves 88% in versus the comparison systems (70% and 78% respectively). Only with the major torso body part at the root of the kinematic tree is the accuracy similar between all three systems (around 95%); head detection is also similar to Ferrari at 90% but improves on Ramanan’s accuracy of 70% for the head. When analysed in terms of the more rigorous pixel error metric, the proposed system also outperforms the other two significantly, with the proposed system averaging 8.5 pixels over all joints in comparison to 13.5 pixels for Ramanan and 12.5 pixels for Ferrari.

The main reasons for the proposed system’s improvement over the system of Ramanan *et al.* (2007) is that the latter only uses a specific (colour-based) appearance model to estimate human pose, ignoring the evidence of a generic appearance model. In addition, Ramanan’s system only uses a single frame to learn the specific appearance model, which affects the robustness of the specific appearance model with respect to lighting variations throughout the video sequence. In contrast, the proposed system uses multiple frames across the sequence to learn a specific (colour-based) appearance model and effectively utilizes both the generic and specific appearance models to achieve the final estimations. In comparison to the approach of Ferrari *et al.* (2009), rather than using the likelihood of generic detections to choose good estimations for learning a specific appearance model,

the more principled method based on clustering provides for an important capability to select accurate generic detections which are then used to learn a more effective specific appearance model.

In Chapter 5, the problem of contamination of the colour appearance model by background pixels is addressed. In order to learn an uncontaminated specific appearance model, an approach is proposed to identify and remove non-target (background) pixels from optimal estimations of the generic human pose detector. Specifically, the characteristics distinguishing target pixels from non-target pixels in optimal estimations are first investigated and analyzed. The results of analysis show that target pixels are more likely to appear on the central area than non-target pixels. Due to the consistency of a body part's appearance in a sequence and the part's difference in colour from non-target pixels, clustering pixels based on colour is applied to separate target pixels into one set of clusters and non-target pixels into other clusters. The cluster with target pixels is identified by selecting the clusters that are largely produced by pixels that are more centrally located in the bounding box. A likelihood model for each body part is then developed and the final estimation for each body part in each frame is derived by choosing the best estimation using the uncontaminated specific appearance feature with the generic detections, similarly to the approach of Chapter 4.

Experiments are conducted to compare the proposed system with Ramanan's system, Ferrari-Core system and the system proposed in Chapter 4. As with Chapter 4, two evaluation metrics (bounding box overlap and pixel error) are again employed. The experimental results show that proposed system improves further upon the approach of Chapter 4, and by implication significantly outperforms the approaches of Ramanan *et al.* (2007) and Ferrari *et al.* (2009). In terms of bounding box overlap, the accuracy of the proposed system in tracking arms exceeds Ramanan, Ferrari-Core and the Chapter 4 systems by 20%, 30% and 6% respectively. Based on the pixel error evaluation, the advantage of the proposed system is more clearly shown, with the average pixel error of the proposed system at 6.3 pixels, compared to 8.6 pixels of the system from Chapter 4. In more details, the most significant improvement by the proposed system is in tracking arms due to the fact that they are typically most difficult to locate correctly and tend to be small, leading to proportionally higher contamination of the arms' colour models. In contrast to the three other systems, the proposed system in this chapter learns a specific appearance model for the limbs that is far less contaminated by the non-target pixels and thus the learned specific appearance model can more clearly detect the limbs, resulting in more accurate estimations, especially for the body part with small size (such as the arms). In tracking arms, the average pixel error of the proposed system is at 6.2 pixels, compared to 9.3 pixels of the system from Chapter 4.

6.1 Summary of Contributions

The contributions of this thesis include the following:

- A quantitative analysis of the relationship between accuracy and likelihood of human pose estimations based on generic detections. We try to find the answers to three questions, which are critical in clearly defining the requirements that a pose estimations system must address in order to achieve robust and accurate tracking:
 - What proportion of optimal per-frame estimations are accurate estimations?
 - Does an accurate estimation consistently exist with the top N estimations of each frame?
 - Is detection likelihood a useful measure for comparing the accuracy of optimal estimations between frames despite being based on different data?
- The use of clustering to learn a robust specific appearance model. The significance of this include:
 - The inclusion of alternative evidence, namely colour, in the process of selecting the accurate generic detections to learn a specific appearance model.
 - This eliminates the need of [Ramanan *et al.* \(2007\)](#) for a special stylized pose to exist in the sequence and will use multiple frames to build the colour model.
 - It avoids the need to rely solely upon the likelihood of the generic detections as [Ferrari *et al.* \(2009\)](#) do, filtering out poor (though high-likelihood) detections for learning the specific appearance model.
- Proposing an effective method to make the final estimations conform to both the generic and specific appearance models. The final estimation is determined by filtering the top few most likely generic estimations based on the specific appearance model. This has several advantages:
 - The evidences from both the generic shaped-based and specific colour-based information are satisfied in the final estimation decision.
 - This is in contrast to [Ramanan *et al.* \(2007\)](#), who discards the generic detection evidence once the specific appearance model is built from it, and [Ferrari *et al.* \(2009\)](#), who averages the two types of evidences.
- The analysis of the characteristics of contamination in accurate estimations, and subsequent proposal of a method to identify and remove non-target (background) pixels from training samples used in learning a specific appearance model. The significance of such an approach is

- a colour model built from a less-contaminated colour profile is less likely to confuse background with the body part to be detected.
- this result in achieving better tracking performance than a colour model that was built with non-target contamination.

6.2 Future Work

In general, there are certain limits to the circumstances that the proposed system can be applied to, regardless of the specific implementation. Foremost is the fact that the proposed system can only be applied in tracking a target whose scale does not change significantly in a sequence. Additionally, motion blur can not be found and processed by the proposed system. Furthermore, the proposed system currently has only been applied to tracking a single person – some adjustments to the clustering process would need to be made in order to track multiple persons. Finally, the efficiency of proposed approach would ideally be improved.

6.2.1 Scale Changes

Tracking a target which experiences large scale changes is a common situation, such as would occur if the scene has a deep field of view and the person moves from the far field to the near field over the course of the sequence. To address this, two possible approaches are as follows. One is to first determine the scale of target in each frame and then build the specific appearance model for that frame. The second is to first build the specific appearance model from similar-scale frames, then determine the scale of target in each frame. The first method to detect the scale of the target could employ the generic human pose detector since such detections can be performed on a range of scales. The idea is to choose the scale with the maximum posterior score as the scale of the target in a frame. In this way, the key to success relies on whether the maximum posterior score will typically represent the best fitting scale. The second approach would be to determine the scale of the target based on the specific appearance model. Specifically, when the specific appearance model is built, one can use it to detect the pixels of target. The number of target pixels in a frame can then be used to provide a rough estimate as to the scale of target in a frame. However, this approach would require a part of the video sequence where the target remains at a similar scale in order to extract enough frames that are similar in scale and so can be used to learn the initial specific appearance model.

6.2.2 Motion Blur

Motion blur, when it occurs, is a problematic factor for bottom-up detection since it not only blurs the edges of a fast-moving body part but also blurs the part’s colour with the background. Thus both generic detectors that use edges and specific detectors based on colour are negatively affected. A possible solution is to identify frames containing significant motion blur. If the specific appearance model is accurate and able to be used to distinguish target pixels from non-target pixels (such as the specific appearance model proposed in Chapter 5), it could be used to find areas where target pixels and non-target pixels coexist and cross. Specifically, given knowledge of the background where the limb is passing over, one could construct a ‘blurred’ colour appearance and verify whether this fits the observed colours. Estimating the exact location of the motion-blurred body part itself would need to be done by using temporal smoothing from surrounding frames, or a spatial search based on other body part to determine the configuration.

6.2.3 Multiple Targets

The proposed system currently does not implement the ability to track multiple people within the same video sequence. However, this could be achieved by extending the proposed system to perform an analysis of the colour clusters, as [Ramanan *et al.* \(2007\)](#) also suggests. In Chapter 4, clustering was used to find the correct appearance of the tracked target. This clustering method can be extended to find the appearance of multiple persons by analyzing what clusters are formed and whether they move. For example, currently in each frame the optimal estimations are clustered to find the largest cluster which is assumed to represent the correct appearance. To extend this to handle multiple people, one could analyze the clustering derived for torso, since this is a large and consistently-detected body part. Specifically, multiple people will produce multiple relatively large clusters of appearance groups. Using the reasoning of [Ramanan *et al.* \(2007\)](#), only those clusters that represent torsos moving within the scene should be considered as possible people to track. In contrast, objects that do not move at all throughout the entire sequence should be ignored as background detections. Once the candidate torso clusters are defined, the other body parts can be analyzed and filtered by selecting the largest groups that attach to these torso candidates and thus are part of the person rather than the background. Note however that issues will still exist when people interact closely together, since the pictorial structure may become confused over which body parts belong to which person.

6.2.4 Efficiency

The main processing load in the proposed systems is in applying the generic human pose detector to obtain the generic detections. This is because the pictorial structure must search every possible position in each frame of the sequence in order to find likely pose configurations. Although this computational cost is a common issue with structure-based bottom-up detectors, some heuristics may be used to reduce the cost. Specifically, since the torso part is at the root of the kinematic tree and relatively large, it may be useful to search for torso parts initially and organize the subsequent search along the kinematic tree. However, a greedy approach may lead to inaccurate detections that violate Chapter 3's findings that the top 30 detections usually contain a correct estimation. Thus any improvement in speed must be made such that the method's basic assumptions remain valid.

Bibliography

- Abd-Almageed, W. and Davis, L. (2007). Robust appearance modeling for pedestrian and vehicle tracking. *Multimodal Technologies for Perception of Humans*, pages 209–215. 115
- Agarwal, A. and Triggs, B. (2006). Recovering 3D human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**(1), 44–58. 22
- Alahi, A., Vandergheynst, P., Bierlaire, M., and Kunt, M. (2010). Cascade of descriptors to detect and track objects across any network of cameras. *Computer Vision and Image Understanding*, **114**(6), 624–640. 17
- Amato, A., Mozerov, M., Bagdanov, A., and Gonzalez, J. (2011). Accurate moving cast shadow suppression based on local color constancy detection. *IEEE Transactions on Image Processing*, **20**(10), 2954–2966. 14
- Andriluka, M., Roth, S., and Schiele, B. (2009). Pictorial structures revisited: People detection and articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1014–1021. 33, 34, 35, 59, 98
- Atsushi, N., Hirokazu, K., Shinsaku, H., and Seiji, I. (2002). Tracking multiple people using distributed vision systems. In *IEEE International Conference on Robotics and Automation*, volume 3, pages 2974–2981. 20
- Bae, J., Hwang, Y., and Choi, B. (2013). Background subtraction using edge cues and color difference for stabilized cmos images. In *2013 IEEE International Conference on Consumer Electronics (ICCE)*, pages 165–166. IEEE. 27
- Balan, A., Sigal, L., and Black, M. (2005). A quantitative evaluation of video-based 3D person tracking. In *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 349–356. 8, 27, 29
- Belongie, S., Malik, J., and Puzicha, J. (2001). Matching shapes. In *IEEE International Conference on Computer Vision*, volume 1, pages 454–461. 19
- Bradski, G. and Davis, J. (2000). Motion segmentation and pose recognition with motion history gradients. In *IEEE Workshop on Applications of Computer Vision*, pages 238–244. 16
- Bregler, C. and Malik, J. (1998). Tracking people with twists and exponential maps. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8–15. 29

- Burak Ozer, I. and Wolf, W. (2002). A hierarchical human detection system in (un)compressed domains. *IEEE Transactions on Multimedia*, **4**(2), 283–300. [17](#)
- Caillette, F., Galata, A., and Howard, T. (2005). Real-time 3-d human body tracking using variable length markov models. In *British Machine Vision Conference*, volume 1, pages 469–478. [27](#)
- Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **8**(6), 679–698. [35](#)
- Carranza, J., Theobalt, C., Magnor, M. A., and Seidel, H.-P. (2003). Free-viewpoint video of human actors. In *ACM SIGGRAPH 2003 Papers*, SIGGRAPH '03, pages 569–577, New York, NY, USA. ACM. [viii](#), [25](#)
- Chari, V., Agrawal, A., Taguchi, Y., and Ramalingam, S. (2012). Convex bricks: A new primitive for visual hull modeling and reconstruction. In *2012 IEEE International Conference on Robotics and Automation (ICRA)*, pages 770–777. IEEE. [27](#)
- Chen, M., Ma, G., and Kee, S. (2005). Pixels classification for moving object extraction. In *IEEE Workshop on Motion and Video Computing*, volume 2, pages 44–49. IEEE. [14](#)
- Chen, Y.-T. and Chen, C.-S. (2008). Fast human detection using a novel boosted cascading structure with meta stages. *IEEE Transactions on Image Processing*, **17**(8), 1452–1464. [16](#)
- Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(5), 603–619. [77](#), [105](#), [115](#)
- Comaniciu, D., Ramesh, V., and Meer, P. (2003). Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(5), 564–577. [18](#)
- Coughlan, J. and Ferreira, S. (2002). Finding deformable shapes using loopy belief propagation. In *European Conference on Computer Vision*, pages 53–73. Springer. [36](#)
- Crandall, D. and Huttenlocher, D. (2006). Weakly supervised learning of part-based spatial models for visual object recognition. In *European Conference on Computer Vision*, pages 16–29. Springer. [31](#), [36](#)
- Crandall, D., Felzenszwalb, P., and Huttenlocher, D. (2005). Spatial priors for part-based recognition using statistical models. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 10–17. [31](#), [36](#)
- Cucchiara, R., Grana, C., Piccardi, M., and Prati, A. (2003). Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(10), 1337–1342. [13](#), [14](#)

- Cucchiara, R., Grana, C., Tardini, G., and Vezzani, R. (2004). Probabilistic people tracking for occlusion handling. In *International Conference on Pattern Recognition*, volume 1, pages 132–135. 18
- Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893. 16, 19
- Dalal, N., Triggs, B., and Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision*, pages 428–441. Springer. 19
- Davis, J. and Bobick, A. (1997). The representation and recognition of human movement using temporal templates. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 928–934. 16
- Davis, L., Philomin, V., and Duraiswami, R. (2000). Tracking humans from a moving platform. In *International Conference on Pattern Recognition*, volume 4, pages 171–178. 20
- Deutscher, J. and Reid, I. (2005). Articulated body motion capture by stochastic search. *International Journal of Computer Vision*, 61, 185–205. 27, 29
- Deutscher, J., Blake, A., and Reid, I. (2000). Articulated body motion capture by annealed particle filtering. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 126–133. 29, 30, 31
- Doyle, D., Jennings, A., and Black, J. (2013). Optical flow background subtraction for real-time ptz camera object tracking. In *2013 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 866–871. 14
- Eichner, M., Marin-Jimenez, M., Zisserman, A., and Ferrari, V. (2012). 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *International journal of computer vision*, 99(2), 190–214. 31
- Elgammal, A., Harwood, D., and Davis, L. (2000). Non-parametric model for background subtraction. In *European Conference on Computer Vision*, pages 751–767. Springer. 13, 14
- Eng, H.-L., Toh, K.-A., Kam, A., Wang, J., and Yau, W.-Y. (2003). An automatic drowning detection surveillance system for challenging outdoor pool environments. In *IEEE International Conference on Computer Vision*, pages 532–539. 13, 14, 15

- Eng, H.-L., Wang, J., Wah, A., and Yau, W.-Y. (2006). Robust human detection within a highly dynamic aquatic environment in real time. *IEEE Transactions on Image Processing*, **15**(6), 1583–1600. [13](#)
- Felzenszwalb, P. and Huttenlocher, D. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision*, **61**(1), 55–79. [2](#), [31](#), [33](#), [36](#), [46](#), [47](#), [49](#), [51](#), [52](#), [54](#), [55](#), [136](#)
- Ferrari, V., Marin-Jimenez, M., and Zisserman, A. (2008). Progressive search space reduction for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. [33](#)
- Ferrari, V., Marín-Jiménez, M., and Zisserman, A. (2009). 2d human pose estimation in tv shows. *Statistical and Geometrical Approaches to Visual Motion Analysis*, pages 128–147. [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [37](#), [39](#), [43](#), [44](#), [45](#), [46](#), [61](#), [62](#), [69](#), [71](#), [72](#), [73](#), [80](#), [82](#), [83](#), [89](#), [91](#), [101](#), [102](#), [103](#), [136](#), [137](#), [138](#), [139](#), [140](#)
- Figueroa, P., Leite, N., and Barros, R. (2006). Background recovering in outdoor image sequences: An example of soccer players segmentation. *Image and Vision Computing*, **24**(4), 363–374. [14](#)
- Fihl, P., Corlin, R., Park, S., Moeslund, T., and Trivedi, M. (2006). Tracking of individuals in very long video sequences. *Advances in Visual Computing*, pages 60–69. [13](#), [14](#), [15](#)
- Fischler, M. and Elschlager, R. (1973). The representation and matching of pictorial structures. *IEEE Transactions on Computers*, **100**(1), 67–92. [46](#)
- Freund, Y. and Schapire, R. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer. [59](#)
- Gao, W., Ai, H., and Lao, S. (2009). Adaptive contour features in oriented granular space for human detection and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1786–1793. [20](#)
- Garcia-Martin, A., Hauptmann, A., and Martinez, J. (2011). People detection based on appearance and motion models. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 256–260. [16](#)
- Geng, L. and Xiao, Z. (2011). Real time foreground-background segmentation using two-layer codebook model. In *International Conference on Control, Automation and Systems Engineering*, pages 1–5. [15](#)

- Gloyer, B., Aghajan, H., Siu, K., and Kailath, T. (1995). Video-based freeway-monitoring system using recursive vehicle tracking. In *Proceedings of SPIE*, volume 2421, page 173. [15](#)
- Gonzalez, J., Lim, I. S., Fua, P., and Thalmann, D. (2003). Robust tracking and segmentation of human motion in an image sequence. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 29–32. [16](#)
- Grauman, K., Shakhnarovich, G., and Darrell, T. (2003). Inferring 3D structure with a statistical image-based shape model. In *International Conference on Computer Vision*, pages 641–647. [21](#), [22](#)
- Guha, P., Mukerjee, A., and Venkatesh, K. (2005). Efficient occlusion handling for multiple agent tracking by reasoning with surveillance event primitives. In *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 49–56. IEEE. [14](#)
- Gutchess, D., Trajkovics, M., Cohen-Solal, E., Lyons, D., and Jain, A. (2001). A background model initialization algorithm for video surveillance. In *IEEE International Conference on Computer Vision*, volume 1, pages 733–740. [15](#)
- Han, B. and Davis, L. (2012). Density-based multifeature background subtraction with support vector machine. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**(5), 1017–1023. [12](#)
- Haritaoglu, I., Harwood, D., and Davis, L. (2000). W4: real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8), 809–830. [13](#), [15](#), [16](#), [20](#)
- Heikkilä, M. and Pietikainen, M. (2006). A texture-based method for modeling the background and detecting moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**(4), 657–662. [13](#)
- Heikkilä, M., Pietikäinen, M., and Heikkilä, J. (2004). A texture-based method for detecting moving objects. In *British Machine Vision Conference*, volume 1, pages 187–196. [13](#)
- Hogg, D. (1983). Model-based vision: a program to see a walking person. *Image and vision computing*, **1**(1), 5–20. [25](#)
- Horprasert, T., Harwood, D., and Davis, L. (1999). A statistical approach for real-time robust background subtraction and shadow detection. In *IEEE International Conference on Computer Vision*, volume 99, pages 256–261. Citeseer. [14](#)

- Hu, M., Hu, W., and Tan, T. (2004). Tracking people through occlusions. In *International Conference on Pattern Recognition*, volume 2, pages 724–727. 18, 19
- Huang, Y. and Huang, T. (2002). Model-based human body tracking. In *International Conference on Pattern Recognition*, volume 1, pages 552–555. viii, 25
- Ilyas, A., Scuturici, M., and Miguet, S. (2009). Real time foreground-background segmentation using a modified codebook model. In *International Conference on Advanced Video and Signal Based Surveillance*, pages 454–459. 15
- Isard, M. and Blake, A. (1998). Condensation – conditional density propagation for visual tracking. *International journal of computer vision*, 29(1), 5–28. 30
- Jain, V. and Crowley, J. L. (2013). Head pose estimation using multi-scale gaussian derivatives. In *Image Analysis*, pages 319–328. Springer. 33
- Jiang, S.-J. F., Muchtar, K., Lin, C.-Y., Kang, L.-W., and Yeh, C.-H. (2012). Background subtraction by modeling pixel and neighborhood information. In *Signal Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–5. 13
- Jin, H. and Qian, G. (2009). People location and orientation tracking in multiple views. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1249–1252. 18
- Kang, J., Cohen, I., and Medioni, G. (2005). Persistent objects tracking across multiple non overlapping cameras. In *IEEE Workshops on Application of Computer Vision*, volume 2, pages 112–119. 18
- Kehl, R. and Gool, L. (2006). Markerless tracking of complex human motions from multiple views. *Computer Vision and Image Understanding*, 104(2), 190–209. viii, 25
- Kim, H., Ku, B., Han, D. K., Kang, S., and Ko, H. (2012). Adaptive selection of model histograms in block-based background subtraction. *Electronics Letters*, 48(8), 434–435. 13
- Kim, K., Chalidabhongse, T., Harwood, D., and Davis, L. (2005). Real-time foreground-background segmentation using codebook model. *Real-time imaging*, 11(3), 172–185. 13, 15
- Koschan, A., Kang, S., Paik, J., Abidi, B., and Abidi, M. (2003). Color active shape models for tracking non-rigid objects. *Pattern Recognition Letters*, 24(11), 1751–1765. 20
- Kristensen, F., Nilsson, P., and Öwall, V. (2006). Background segmentation beyond rgb. In *Asian Conference on Computer Vision*, pages 602–612. Springer. 12

- Krüger, V., Anderson, J., and Prehn, T. (2005). Probabilistic model-based background subtraction. *Image Analysis*, pages 567–576. [20](#)
- Lan, X. and Huttenlocher, D. (2005). Beyond trees: common-factor models for 2D human pose recovery. In *International Conference on Computer Vision*, volume 1, pages 470–477. [31](#), [36](#)
- Leibe, B., Seemann, E., and Schiele, B. (2005). Pedestrian detection in crowded scenes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 878–885. [19](#)
- Li, C., Guo, L., and Hu, Y. (2010a). A new method combining HOG and Kalman filter for video-based human detection and tracking. In *International Conference on Image and Signal Processing*, volume 1, pages 290–293. [20](#)
- Li, Q., Shao, C., Yue, H., and Li, J. (2010b). Real-time foreground-background segmentation based on improved codebook model. In *International Congress on Image and Signal Processing*, volume 1, pages 269–273. [15](#)
- Ligorio, G. and Sabatini, A. M. (2013). Extended kalman filter-based methods for pose estimation using visual, inertial and magnetic sensors: Comparative analysis and performance evaluation. *Sensors*, **13**(2), 1919–1941. [29](#)
- Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, **60**(2), 91–110. [19](#)
- McKenna, S., Jabri, S., Duric, Z., and Wechsler, H. (2000). Tracking interacting people. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 348–353. [13](#), [14](#), [18](#)
- Mikic, I., Trivedi, M., Hunter, E., and Cosman, P. (2001). Articulated body posture estimation from multi-camera voxel data. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–455. IEEE. [27](#)
- Mikolajczyk, K. and Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**(10), 1615–1630. [33](#), [35](#), [59](#)
- Ming-yu, C. and Hauptmann, A. (2009). *MoSIFT: Recognizing human actions in surveillance videos*. Ph.D. thesis, Pittsburgh: Carnegie-Mellon University. [16](#)
- Mittal, A. and Davis, L. (2003). M2 tracker: a multi-view approach to segmenting and tracking people in a cluttered scene. *International Journal of Computer Vision*, **51**(3), 189–203. [18](#)

- Mohan, A., Papageorgiou, C., and Poggio, T. (2001). Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**(4), 349–361. 17
- Monnet, A., Mittal, A., Paragios, N., and Ramesh, V. (2003). Background modeling and subtraction of dynamic scenes. In *International Conference on Computer Vision*, pages 1305–1312. 13, 14
- Mori, G. and Malik, J. (2006). Recovering 3D human body configurations using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**(7), 1052–1062. 23
- Mori, G., Ren, X., Efros, A., and Malik, J. (2004). Recovering human body configurations: combining segmentation and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 31, 32
- Munder, S., Schnorr, C., and Gavrila, D. (2008). Pedestrian detection and tracking using a mixture of view-based shape–texture models. *IEEE Transactions on Intelligent Transportation Systems*, **9**(2), 333–343. 20
- Okuma, K., Taleghani, A., Freitas, N., Little, J., and Lowe, D. (2004). A boosted particle filter: Multitarget detection and tracking. In *European Conference on Computer Vision*, pages 28–39. Springer. 17, 18, 19
- Oliver, N., Rosario, B., and Pentland, A. (2000). A Bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8), 831–843. 13
- O’Rourke, J. and Badler, N. (1980). Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **2**(6), 522–536. 25
- Park, S. and Aggarwal, J. (2006). Simultaneous tracking of multiple body parts of interacting persons. *Computer Vision and Image Understanding*, **102**(1), 1–21. 19
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann. 47
- Peursum, P., Venkatesh, S., and West, G. (2007). Tracking-as-recognition for articulated full-body human motion analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE CS Press. 27
- Peursum, P., Venkatesh, S., and West, G. (2010). A study on smoothing for particle-filtered human body tracking. *International Journal of Computer Vision, Special Issue*

- on *Evaluation of Articulated Human Motion and Pose Estimation*, **87**(1/2), 53–74. [29](#), [31](#)
- Porikli, F., Tuzel, O., and Meer, P. (2006). Covariance tracking using model update based on lie algebra. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 728–735. [17](#)
- Prati, A., Mikic, I., Trivedi, M., and Cucchiara, R. (2003). Detecting moving shadows: algorithms and evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(7), 918–923. [12](#)
- Ramanan, D. (2006). Learning to parse images of articulated bodies. *Advances in neural information processing systems*, **19**, 1129. [39](#)
- Ramanan, D., Forsyth, D., and Zisserman, A. (2007). Tracking people by learning their appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29**(1), 65–81. [viii](#), [ix](#), [x](#), [2](#), [3](#), [4](#), [5](#), [7](#), [8](#), [9](#), [37](#), [38](#), [40](#), [41](#), [42](#), [43](#), [44](#), [45](#), [61](#), [62](#), [63](#), [71](#), [72](#), [81](#), [82](#), [89](#), [90](#), [91](#), [92](#), [93](#), [96](#), [101](#), [102](#), [103](#), [125](#), [136](#), [137](#), [138](#), [139](#), [140](#), [142](#)
- Ren, X., Berg, A., and Malik, J. (2005). Recovering human body configurations using pairwise constraints between parts. In *IEEE International Conference on Computer Vision*, volume 1, pages 824–831. [31](#), [32](#)
- Rohr, K. (1994). Towards model-based recognition of human movements in image sequences. *CVGIP-Image Understanding*, **59**(1), 94–115. [25](#)
- Ronfard, R., Schmid, C., and Triggs, B. (2002). Learning to parse pictures of people. In *European Conference on Computer Vision*, pages 700–714. [33](#), [34](#)
- Sabzmeydani, P. and Mori, G. (2007). Detecting pedestrians by learning shapelet features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. [19](#)
- Sangi, P., Heikkilä, J., and Silvén, O. (2001). Extracting motion components from image sequences using particle filters. In *PROCEEDINGS OF THE SCANDINAVIAN CONFERENCE ON IMAGE ANALYSIS*, pages 508–514. [16](#)
- Seemann, E., Leibe, B., Mikolajczyk, K., and Schiele, B. (2005). An evaluation of local shape-based features for pedestrian detection. In *BMVC*. Citeseer. [35](#), [59](#)
- Sheikh, Y. and Shah, M. (2005). Bayesian modeling of dynamic scenes for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**(11), 1778–1792. [14](#)
- Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., *et al.* (2013a). Efficient human pose estimation

- from single depth images. In *Decision Forests for Computer Vision and Medical Image Analysis*, pages 175–192. Springer. 33
- Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., and Moore, R. (2013b). Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1), 116–124. 31
- Sidenbladh, H. (2004). Detecting human motion with support vector machines. In *International Conference on Pattern Recognition*, volume 2, pages 188–191. 16
- Sidenbladh, H., Black, M., and Fleet, D. (2000). Stochastic tracking of 3D human figures using 2D image motion. In *European Conference on Computer Vision*, pages 702–718. Springer. viii, 25, 26, 30
- Sigal, L. and Black, M. (2006). Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Brown University TR*, 120. x, xi, 89, 91, 108, 109, 110, 113, 114, 120, 125
- Sigal, L., Isard, M., Sigelman, B. H., and Black, M. J. (2003). Attractive people: Assembling loose-limbed models using non-parametric belief propagation. In *NIPS*, pages 1539–1546. MIT Press. 31, 36
- Sminchisescu, C. and Jepson, A. (2004). Variational mixture smoothing for non-linear dynamical systems. In *IEEE Conference on Computer Vision and Pattern Recognition*. 27
- Sminchisescu, C., Kanaujia, A., Li, Z., and Metaxas, D. (2005). Discriminative density propagation for 3D human motion estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 390–397. 23
- Stauffer, C. and Grimson, W. (1999). Adaptive background mixture models for real-time tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 10, 12, 13, 14
- Taylor, C. (2000). Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 677–684. 18, 19
- Toyama, K. and Blake, A. (2002). Probabilistic tracking with exemplars in a metric space. *International Journal of Computer Vision*, 48(1), 9–19. 58
- Tuzel, O., Porikli, F., and Meer, P. (2006). Region covariance: A fast descriptor for detection and classification. In *European Conference on Computer Vision*, pages 589–600. Springer. 17

- Tuzel, O., Porikli, F., and Meer, P. (2007). Human detection via classification on riemannian manifolds. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. 17
- Urtasun, R. and Darrell, T. (2008). Sparse probabilistic regression for activity-independent human pose inference. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. 23
- Utsumi, A. and Tetsutani, N. (2002). Human detection using geometrical pixel value structures. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 34–39. 17
- Vezzani, R., Grana, C., and Cucchiara, R. (2011). Probabilistic people tracking with appearance models and occlusion classification: The AD-HOC system. *Pattern Recognition Letters*, 32(6), 867–877. 18
- Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 511–518. 16
- Viola, P., Jones, M., and Snow, D. (2003). Detecting pedestrians using patterns of motion and appearance. In *IEEE International Conference on Computer Vision*, pages 734–741. 16
- Wachter, S. and Nagel, H.-H. (1997). Tracking of persons in monocular image sequences. In *Nonrigid and Articulated Motion Workshop*, pages 2–9. 25
- Wang, H. and Suter, D. (2005). Background initialization with a new robust statistical approach. In *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 153–159. 15
- Wang, H. and Suter, D. (2006). A novel robust statistical method for background initialization and visual surveillance. In *Asian Conference on Computer Vision*, pages 328–337. Springer. 15
- Wells, W. M. (1986). Efficient synthesis of gaussian filters by cascaded uniform filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(2), 234–239. 56
- Wren, C., Azarbayejani, A., Darrell, T., and Pentland, A. (1997). Pfunder: real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7), 780–785. 13

- Wu, B. and Nevatia, R. (2005). Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *IEEE International Conference on Computer Vision*, volume 1, pages 90–97. 21
- Wu, Y. and Yu, T. (2006). A field model for human detection and tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5), 753–765. 19
- Xu, L.-Q. and Puig, P. (2005). A hybrid blob- and appearance-based framework for multi-object tracking through complex occlusions. In *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 73–80. 18
- Xu, S. (2009). Dynamic background modeling for foreground segmentation. In *Computer and Information Science*, pages 599–604. 14
- Yang, C., Duraiswami, R., and Davis, L. (2005a). Fast multiple object tracking via a hierarchical particle filter. In *IEEE International Conference on Computer Vision*, volume 1, pages 212–219. 18, 19
- Yang, T., Pan, Q., Li, J., and Li, S. (2005b). Real-time multiple objects tracking with occlusion handling in dynamic scenes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 970–975. 14
- Yang, Y. and Chen, W. (2012). Parallel algorithm for moving foreground detection in dynamic background. In *2012 Fifth International Symposium on Computational Intelligence and Design (ISCID)*, volume 2, pages 442–445. 14
- Yin, Z. and Collins, R. (2006). Moving object localization in thermal imagery by forward-backward MHI. In *Computer Vision and Pattern Recognition Workshop*, page 133. 16
- Zhang, S., Yao, H., and Liu, S. (2008). Dynamic background modeling and subtraction using spatio-temporal local binary patterns. In *IEEE International Conference on Image Processing*, pages 1556–1559. 14
- Zhao, L. and Davis, L. (2005). Closely coupled object detection and segmentation. In *IEEE International Conference on Computer Vision*, volume 1, pages 454–461. 19
- Zhao, T. and Nevatia, R. (2004a). Tracking multiple humans in complex situations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9), 1208–1221. 14, 18
- Zhao, T. and Nevatia, R. (2004b). Tracking multiple humans in crowded environment. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 406–413. 13, 18

- Zhao, T., Nevatia, R., and Wu, B. (2008). Segmentation and tracking of multiple humans in crowded environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**(7), 1198–1211. [18](#)
- Zhong, J. and Sclaroff, S. (2003). Segmenting foreground objects from a dynamic textured background via a robust Kalman filter. In *International Conference on Computer Vision*, pages 44–50. [13](#), [14](#)
- Zhu, Q., Yeh, M.-C., Cheng, K.-T., and Avidan, S. (2006). Fast human detection using a cascade of histograms of oriented gradients. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1491–1498. [16](#)